DOCUMENT RESUME

ED 135 797 TM 004 877

AUTHOR Paisley, Matilda B.; And Others
TITLE The Convergent-Discriminant Matrix: Multitrait-Multimethod Logic Extended to Other Social Research Decisions.
PUB DATE Sep 70
NOTE 162p.

EDRS PRICE MF-$0.83 HC-$8.69 Plus Postage.
DESCRIPTORS Correlation; Matrices; *Measurement Techniques; Reliability; *Research Methodology; Sampling; Validity
IDENTIFIERS *Convergent Discriminant Matrices; *Multitrait Multimethod Techniques

ABSTRACT

Multitrait-multimethod methodology consists of measuring several different traits on a particular sample using several different measurement techniques or settings (methods) with each trait. The purpose of the methodology is to simultaneously show that different methods of measuring a particular trait correlate highly (converge), and that different traits measured with the same method have small correlations with the trait of interest (discriminate). This paper reviews the multitrait-multimethod logic, surveys uses to which multitrait-multimethod methodology has been put, and extends the multitrait-multimethod argument to that of a more general "convergent-discriminant comparison" which goes beyond methods and traits. Five classes of research decisions are proposed for use with the convergent-discriminant methodology: concepts, measures, populations, times, and analysis models. The ten possible resulting matrices (from combinations of the five classes, two at a time) are described in detail and technical issues related to them are discussed. Finally, convergent-discriminant methodology is related to various causal analysis techniques drawn from the literature. These techniques include the sixteen-fold table, cross-lagged correlation, frequency in change of product-moment, and path analysis. (JKS)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED135797

THE CONVERGENT-DISCRIMINANT MATRIX: MULTITRAIT-MULTIMETHOD LOGIC EXTENDED TO OTHER SOCIAL RESEARCH DECISIONS

Matilda B. Paisley
Department of Psychology
Stanford University

W. Andrew Collins
William J. Paisley
Institute for Communication Research
Stanford University

September 1970

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM004 877

FOREWORD

We have been sitting on parts of this manuscript since 1968. As we pursued extensions of the multitrait-multimethod matrix, each question that we "answered" seemed to raise two more. If we didn't find convergent-discriminant thinking so useful in our own research, we would set the manuscript aside until its loose ends could be gathered up. However, we suspect that loose ends are in the nature of the problem.

A project of this kind accrues many debts -- for example, to faculty and students at Stanford who made suggestions and contributed data sets. We thank everyone who helped us, and we hope that the paper now is useful to them either in research or in teaching.

We must also acknowledge our intellectual debt to Donald T. Campbell and Donald W. Fiske for "Convergent and discriminant validation by the multitrait-multimethod matrix" (1959) and to Raymond B. Cattell for "The three basic factor-analytic designs -- their interrelations and derivatives" (1952). The first article planted the thought that convergent-discriminant logic might be extended beyond method-trait comparisons. The second article suggested other correlatable dimensions on which tests of convergence and discrimination would make research decisions more tough-minded.

Comments and criticisms will be appreciated.

M.B.P.
W.A.C.
W.J.P.

CONTENTS

Foreword i

I. THE CONCEPTUAL FRAMEWORK

- I.1. The principle of multiple operationism . I/1
- I.2. The multitrait-multimethod matrix . . I/2
- I.3. Taking a broader view of multiple operationism . . I/5
- I.4. Five research choice-points . . . I/6
  - Concepts I/9
  - Measures I/10
  - Populations I/11
  - Times I/12
  - Analysis models I/13
- I.5. The power of single variations . . I/14
- I.6. Double variations and the five research choice-points . . I/16
- I.7. Convergent versus discriminant uses for the matrix . I/18

II. TECHNICAL ISSUES

- II.1. The need for a third, or replication, dimension . II/1
- II.2. Correlating independent data sets . II/2
  - Profile correlation . . . II/3
  - Aggregate correlation . . . II/7

A caveat concerning aggregate correlation . . II/8

A second caveat II/14

II.3. Choosing a correlation coefficient . II/16

II.4. A terminology for discussing the convergent-discriminant matrix . II/17

III. THE TEN CONVERGENT-DISCRIMINANT MATRICES

III.1. Concepts by measures III/2

III.2. Concepts by populations . . . III/20

III.3. Concepts by times III/26

III.4. Concepts by analysis models . . III/34

III.5. Measures by populations . . . III/39

III.6. Measures by times III/42

III.7. Measures by analysis models . . III/45

III.8. Populations by times III/48

III.9. Populations by analysis models . . III/52

III.10. Times by analysis models . . . III/60

IV. THE ROLE OF THE CONVERGENT-DISCRIMINANT MATRIX IN CAUSAL ANALYSIS: FORETHOUGHTS

IV.1. The language of causality . . . IV/2

IV.2. The techniques IV/10

IV.2.A. Sixteenfold table IV/11

IV.2.B. Cross-lagged panel correlation . . IV/14

IV.2.C. Frequency of change in product moment . . IV/23

IV.2.D. Heise's path analysis . . . IV/26
IV.2.E. A comparison of the techniques . . IV/27
IV.2.F. Spurious relationships . . . IV/30
IV.2.G. Data analysis IV/32
IV.2.H. Data inspection in the pre-analysis stage . IV/32
IV.2.I. Data procedures in the mid-analysis stage . IV/33
IV.2.J. Data procedures in the post-analysis stage . IV/35
IV.2.K. Overview IV/37

V. ON THE CORRELATION COEFFICIENT

Correlation versus individual fluctuation V/1

VI. SOME INDICATED DIRECTIONS FOR RESEARCH

Use of the CDM throughout the history of a research project . VI/1

References R/1

THE CONVERGENT-DISCRIMINANT MATRIX: MULTITRAIT-MULTIMETHOD LOGIC EXTENDED TO OTHER RESEARCH DECISIONS

I. THE CONCEPTUAL FRAMEWORK

I.1. The Principle of Multiple Operationism

The product of science is reliable knowledge. The phrase "reliable knowledge" subsumes the reliability and validity criteria of research adequacy: reliable if the findings are stable and consistent; valid if what we know, thanks to the findings, corresponds to what we think we know.

Reliable knowledge sometimes comes through single measurements. More often, especially in the social sciences, multiple measurements or even multiple kinds of measurements are needed.

We are not concerned here with the dispute between single operationism and multiple operationism. Some concepts are properly associated with a single measure or operation; others are properly associated with multiple measures. Whereas "age" and "length" probably allow single measurement, concepts such as "intelligence" and "ethnocentrism" seem to require a triangulation (Campbell, 1956) or convergent validation (Campbell and Fiske, 1959) via multiple, independent measures. A fruitful paradigm for testing convergence is the Campbell-Fiske multitrait-multimethod matrix (1959).

This paper:

1. Reviews multitrait-multimethod logic;
2. Surveys uses to which the "M^3" has been put;
3. Generalizes the Campbell-Fiske argument of convergent-discriminant comparison beyond methods and traits in a convergent-discriminant matrix;
4. Suggests five classes of research decisions or choice-points;
5. Considers application of a convergent-discriminant matrix to them.

I.2. The Multitrait-Multimethod Matrix

Campbell and Fiske (1959) argue that, while method convergence alone is often taken as validation of a concept, the discriminating power of methods must also be demonstrable. Not only should different measures of the same trait correlate highly with each other, they should correlate near zero with measures of other traits -- to the extent that the other traits are conceptually independent of the first. Strong correlations between measures of different traits are suspect as artifactual, especially if the measures share a method factor, such as a common mode of administration (e.g., paper-and-pencil test) or a response set (e.g., acquiescence).

After stating the case for validation that balances convergence with discrimination, Campbell and Fiske propose the multitrait-multimethod matrix for testing both tendencies among measures. Instead of presenting again their original synthetic "M^3" (for which the reader is urged to consult the 1959 article), we present in Table I.2/1 a new and unpublished matrix (Rees and Campbell, 1968) which shows correlations between two measures (own response and socially desirable response) of five traits (rebellion, female careerism, striving, labor, and skepticism).

Table I.2/1. A Specimen Multitrait-Multimethod Matrix.

		Own Response					Socially Desirable Response				
		R	C	S	L	Sk	R	C	S	L	Sk
Own Response	R	(86)									
	C	34	(84)								
	S	11	10	(68)							
	L	15	10	03	(84)						
	Sk	41	30	-01	-17	(71)					
Socially Desirable Response	R	*46*	17	04	18	-01	(85)				
	C	19	*40*	07	12	-06	37	(91)			
	S	-04	-03	*57*	-10	-11	14	-04	(73)		
	L	26	17	03	*29*	01	08	24	-22	(84)	
	Sk	19	06	00	05	*18*	41	16	14	01	(70)

Describing the table in Campbell-Fiske terms, we find two kinds of blocks. The *monomethod* block, one for each method, consists of a *reliability* diagonal of monotrait-monomethod values (parenthesized in this example) and the adjoining triangle of heterotrait-monomethod correlations. The *heteromethod* block consists of a *validity* diagonal of monotrait-heteromethod values (underlined) and two adjoining triangles of heterotrait-heteromethod correlations.

The following questions are asked when examining a multitrait-multimethod matrix:

1. Are the validity values significant? While those in Table I.2/1 seem fairly high, two (L and Sk) are only marginally significant.
2. Is each validity value higher than the values lying in its vectors in the heteromethod block? In Table I.2/1, all but Sk meet this requirement.
3. Does the same trait measured by different methods correlate higher than different traits measured by the same method? This condition is met by three of the five traits. The R-Sk correlations are too high in both monomethod blocks, as is the C-Sk correlation in one monomethod block.
4. Does the same pattern of trait interrelationships emerge in the heterotrait triangles for both the monomethod and heteromethod blocks? In Table I.2/1 the two monomethod blocks have similar patterns of relationships. However, the heteromethod block shows a different pattern in its two triangles, lacking, for example, the strong R-Sk correlations.

Ideally, the multitrait-multimethod matrix would show complete independence for both methods and traits, if methods and traits were indeed chosen as independent. Proof of this would be zero correlations in the monomethod and heteromethod triangles. More often, the best that can be hoped for is a test of relative validity: "The common variance specific to a trait, above and beyond shared method variance." (Campbell and Fiske, 1959, p. 84). Table I.2/1 shows a lack of independence between the two methods, compounded by a lack of independence among the six traits. Particularly, rebellion and skepticism seem to be on the same dimension.

However, each researcher looks at the matrix with some knowledge of his concepts' relationships, and a monomethod correlation between two traits that is *almost* as high as the validity value of each might be acceptable.

I.3. Taking a Broader View of Multiple Operationism

If we think of multiple operationism primarily as *method variation*, with additional concepts introduced only to show that methods discriminate as well as converge, then we do not extend the principle as far as it should be extended in behavioral science. *Method* is only one of several research choice-points at which we should prefer *deliberate variation* as a way of increasing confidence in our findings. Deliberate variation can also be introduced in the set of *concepts* themselves, in the *populations* to be studied, in the *times* of measurement, and in statistical *models* chosen for the analysis.

In this paper we define multiple operationism as *a principle of deliberate variation at every research choice-point where a reliability or validity challenge exists*. Figure I.3/1 illustrates our position.

Some research strategies here gathered under a broader multiple operationism are familiar to behavioral scientists under other names, such as test-retest reliability, validation across independent populations, and the cross-lagged correlation. In those cases, we want to draw out relations among the strategies that may suggest additional applications. Other strategies, such as deliberate variation of analysis model in a convergent-discriminant matrix, seem new and untested.

I.4. Five Research Choice-points

A decision-tree analysis of the research process shows dozens of points, or nodes, at which the researcher may choose a single path or multiple paths for his project to follow. Paisley (1967) discussed 25 "decision points in survey research" in a brief treatment that excluded micro-decisions, such as the exact construction of an instrument. Figure I.4/1 shows a more generic list of choice-points (from Paisley, in preparation).

There are other such outlines in the literature. Any one of them can be made arbitrarily long by including micro-decisions. At a higher level of abstraction, we choose to aggregate all decisions in just five classes, which we label concepts, measures, populations, times and analysis models. A parallel with Cattell's (1952) three dimensions of variables, persons, and times is clear. We differ in distinguishing concepts from measures within the broad class of variables and in adding the dimension of analysis model.[1]

---

[1] Cattell (1966) expanded the variables-persons-times data cube into a ten-dimensional "Basic Data Relation Matrix" encompassing states of stimuli, responses, persons (subjects), observers, and the background ("ambient stimuli"). This impressive system will certainly win a following, especially among experimental psychologists. However, because of our interest in quasi-experimental and non-experimental research, we find that Cattell's earlier

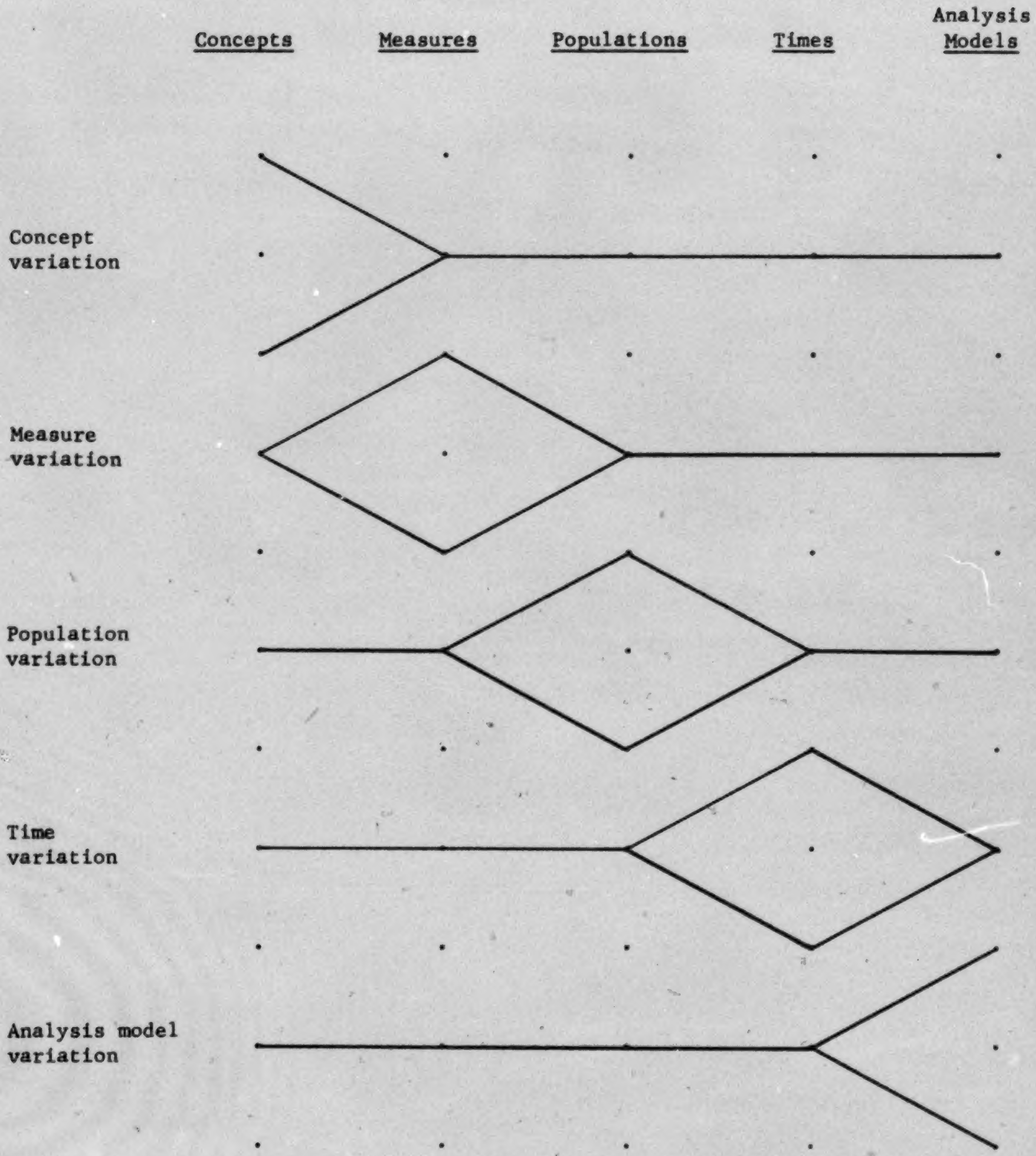

Concepts
Measures
Populations
Times
Analysis Models
Concept variation
Measure variation
Population variation
Time variation
Analysis model variation

Figure I.3/1. Multiple operationism -- as a principle of deliberate variation -- extends to all research choice-points.

Figure I.4/1. CORA paradigm of the research process. Illustrative choice-points within Conceptualization, Observation, Reduction, and Analysis phases.

Conceptualization:
C1. Specification of phenomenon of interest.
C2. Theoretical frame of reference.
C3. Mode of inquiry (exploratory, hypothesis-testing, etc.)
C4. Conceptual definitions.
C5. Operational definitions.
C6. Desired inference.

Observation:
01. Specification of universe of objects to be studied.
02. Design (experimental, quasi-experimental, non-experimental, etc.)
03. Sampling procedure.
04. Occasions of measurement.
05. Instruments of measurement (completion of C5).
06. Mechanical arrangements for measurement.
07. Reliability/validity checks.

Reduction:
R1. Unitization of continuous observational content, if any.
R2. Categorization rules (operational definitions).
R3. Assignment of units to categories (repeat 07).

Analysis:
A1. Goal in analysis (description, explanation, etc.)
A2. Mechanical arrangements for analysis.
A3. Statistics of difference and association.
A4. Operational conclusion.
A5. Conceptual conclusions (inferences of internal validity and external validity).

Because these five classes encompass so many decisions, it is difficult to define them rigorously. Boundaries between concepts and measures, between measures and analysis models, and even between concepts and populations are not as clear as they seem.

Concepts. As Kerlinger says, a concept is "a word that expresses an abstraction formed by generalization from particulars." (1964, p. 4). We would add that a concept, if scientifically fruitful, is related to other concepts as are "the knots in a network of systematic interrelationships in which laws and theoretical principles form the threads." (Hempel, 1966, p. 94). These might be called the property definition and relational definition of a concept, respectively. A functional or pragmatic definition (ours) might be: "a way of talking about a phenomenon, so that past, present, or future behavior is accounted for and relationships with other phenomena are clarified."

The set of concepts to be studied depends, of course, on one's research goal. More concepts are included in a theory-building

---

1 (cont.) data cube is a better springboard for our ideas. The special character of the "Basic Data Relation Matrix" can be seen in the ten dimensions (Cattell, 1966, p. 86): (1) focal stimulus; (2) variant on stimulus; (3) acts, responses, or behavior process; (4) stage of behavior process; (5) persons (or organisms); (6) states (modulators, roles, learning); (7) environmental backgrounds (ambient stimuli); (8) phase of environmental background; (9) observer; (10) condition of observer.

study than in a theory-testing study. When a concept is easily operationalized (e.g., "age"), it is less necessary to include related concepts showing patterns of intercorrelation than in the case, for example, of intelligence, creativity, verbal facility, and other concepts from that domain.

We distinguish concept from measure in the spirit of Hempel's many-to-one argument (1952) -- that each measure serves as a partial interpretation of a concept without providing its complete definition. Many measures may interpret the same concept, and one measure may interpret more than one concept. An example of the latter: a long problem-solving test will reflect not only on intelligence but also on perseverance.

There is much independence, therefore, between the choice of concepts to be studied and the choice of measures to provide data on them.

The power of the convergent-discriminant matrix is exploited when chosen concepts are theoretically linked (e.g., "frustration" and "aggression"), when they present a definitional problem (e.g., "intelligence" and "creativity"), or when their known empirical relationship serves as a touchstone for comparing measures, populations, times, or analysis models.

Measures. Minimally, a measure is a statement about a phenomenon, locating the phenomenon on one or more attribute dimensions. In terms of research purpose, a measure consists of activities that assign empirical meaning to a concept (adapted from Kerlinger, 1964, p. 34). Measures may involve instrumentation or the unaided human senses. Measures may yield quantitative or qualitative data.

Measures may be part of an "obtrusion" upon a natural phenomenon, as when the experimenter manipulates and controls, or they may be an unobtrusive monitoring of the phenomenon *in situ*.

As one of the five classes of research choice-points, measures in our usage must encompass more than measurement alone. We include in this class *all operations required to fill a data matrix*. Therefore the experimenter's manipulations, although not technically measurements, belong to this class. For example, in the Borgatta study discussed in III.1, the group settings (manipulated social context) are not themselves part of the measurement process, but we consider "self-rankings within an intimate group" and "self-rankings within a stratified group" to be variations within the class of measures, even though, narrowly construed, the measure in each variation is the same. The advantage for certain purposes of Cattell's distinctions (1966) is here obvious.

We also include in the measures class the post-measurement, pre-analysis phase of content analysis, or "the transformation of unstructured observational content into data that can be summarized and compared." (Paisley, in preparation). Other post-measurement steps, such as estimation of missing data and computation of index scores, are included in the measures class either if they are required to fill a data matrix *or* if (as in the case of index scores) they are specified in the operational definition of a concept.

Populations. This class consists of the units or objects whose attributes are measured. The conventional scientific meaning of "population" fits this usage, provided that the singular form, "person" or "object," is also available. That is, comparisons are sometimes made between entire populations in the social-political sense, sometimes between persons, and sometimes between measurements of a single person at different points in time or in different roles. Thus, the convergent-discriminant matrix might be used to compare, on some conceptual relationship, the U.S. with the U.S.S.R., or Smith with Jones, or Smith in 1960 with Smith in 1970 (or, for that matter, the U.S. in 1960 with the U.S.S.R. in 1970, and vice versa, in the *populations by times* matrix).

We would not expect the population class to overlap with other classes. That is, concepts are concepts, measures are measures, objects are objects, etc. However, in Q methodology (Stephenson, 1953, *inter alia*), "Jones as he is," "Jones as he remembers himself in childhood," and "Jones as he would like to be" are both *objects to be measured* and *concepts to be related and interpreted*. Such boundary problems can be tolerated if we remember the arbitrariness of any set of classes.

Times. Not even time is a completely obvious class. Time refers to occasion of measurement, and its meaning is usually clear. However (again in a kind of Q or profile analysis), we can think of "the spirit of the 90's," "the spirit of the 20's," and "the spirit of the 60's" in comparing decades that may have more in common with each other than with contiguous decades. Here the object or population is assumed to be constant (e.g., the United States at three points in time) while time is both an *occasion of measurement* and a *concept to be related and interpreted*. Technically, in Cattell's usage, this is not Q but O analysis -- the correlation of occasions, summed over variables.

Analysis models. To the experimental behavioral scientist this will not seem to be a significant choice-point. Following Cronbach's distinction between experimental and correlational research (1957), the analysis model is foreordained in a well-designed experiment. For example, an experiment that hypothesizes mean differences and varies two levels within each of three conditions is a 2x2x2 factorial design whose analysis model will be a 2x2x2 ANOVA if the dependent variable has an interval metric and if other ANOVA assumptions are met. Other dependent variable metrics similarly dictate other analysis models.

However, in non-experimental or correlational research, usually neither design nor data dictate what the analysis model shall be. Survey data typically include interval, ordinal, and nominal metrics in variable sets that we would like to combine in a single analysis model. No matter which analysis model is chosen for such mixed-metric data, it is likely that some variables will be treated less powerfully than their metrics permit while others will violate assumptions governing use of the model.

Faced with this dilemma, the researcher who analyzes mixed-metric data usually relies upon face validation and upon the field's consensus of which assumptions he can disregard. There are only spotty comparative data of the effects of using alternative statistics (e.g., Carroll, 1961; Paisley and Ruggels, 1965).

The mixed-metric problem may be one of the most frustrating that a non-experimental researcher faces, but the necessity of choosing an analysis model without explicit guidance from design or data is inherent in non-experimental research. Even if data have interval metrics and normal distributions, considerations beyond the design and the data suggest linear versus curvilinear regression, orthogonal versus oblique factor analysis, a three-variable versus four-variable causal model, etc.

With the advent of computing power and sophisticated multivariate procedures, with the rigor brought to non-experimental research by such methodologists as Campbell and Stanley (1963), and with the emerging policy significance of non-experimental research on social problems, it is time to improve upon face validation and consensus as criteria for choosing analysis models. With the convergent-discriminant matrix, we can compare outcomes of two or more analysis models, across concepts, measures, populations, or times.

I.5. The Power of Single Variations

Cronbach and Meehl (1955) define construct validity in terms of the relationship between a method and the construct it is supposed to measure. The construct is not available in the form of an external criterion, as in the case of concurrent or predictive validity. The construct is inferred at the core of variance shared by multiple methods thought to measure it.

In the Cronbach-Meehl view, methods reflect the unmeasurable -- the construct -- through their intercorrelations. Thus, for example, alternative measures of "anxiety" are correlated, the weak correlates discarded, and the strong correlates retained in a composite measure with some degree of construct validity. Construct validation differs from item analysis in its emphasis on different, rather than maximally similar measures of the construct. In both cases, replication is needed to check shrinkage in correlations when measure-construct relationships are generalized beyond the original data set.

When we compare multiple measures, we are varying only one of the five dimensions shown in Figure I.4/1. The other four dimensions are assumed either to be invariant or to be changing in irrelevant ways (e.g., a subject can complete only one test at one time; therefore time changes slightly from the beginning of one test to the beginning of the next). We might call this the strategy of single variations.

The strategy of single variations gives us all our measures of reliability. It also provides for time-series analysis of a single concept (more than one concept would constitute a second dimension of variation), replication of a single relationship across populations, the study of many concepts in a single population, etc. In short, the strategy of single variations accounts for the great majority of all past and present research.

What are the limitations of the strategy of single variations? Building upon the strong foundation of Cronbach and Meehl, Campbell and Fiske (1959) argued that single variations were insufficient for construct validation. This argument was summarized in section I.2 above -- briefly: convergence of measures is not enough to triangulate a concept; the measures in question might have a tendency to converge simply because correlational artifacts link them. Several paper-and-pencil tests might share this property. So might tests lacking correction for an acquiescence artifact. The essence of the Campbell-Fiske argument for the multitrait-multimethod matrix is that a second dimension of variation -- namely, concepts -- must be introduced to show that the methods are able to discriminate when discrimination is expected.

Therefore, although the strategy of single variations establishes reliability and provides most of the substantive contrasts in our research literature, Campbell and Fiske turn attention to the strategy of double variations. "Double variationism" is a strategy of great power both within and beyond the domain of concepts and measures in which Campbell and Fiske first applied it.

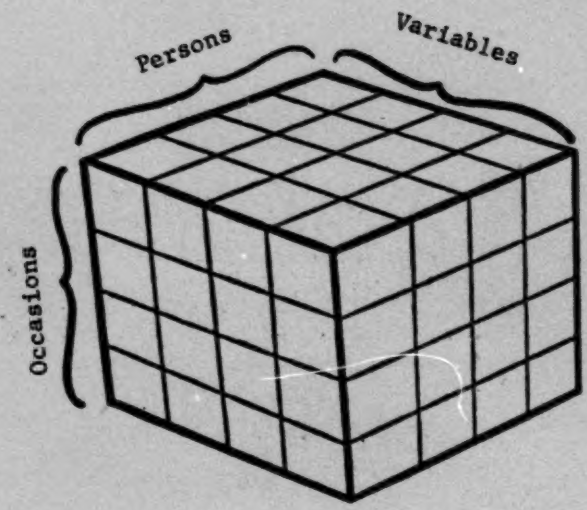
I.6. Double variations and the Five Research Choice-points

We extend the principle of multiple operationism to any deliberate variation of concepts, measures, populations, times, and analysis models. Pursuing the strategy of double variations, we have ten pairs of dimensions, jointly varied, to discuss.

This discussion will be reminiscent of Cattell's review of O, P, Q, R, S, and T techniques (1952), in which it was shown that a data cube of persons, variables, and times permits six distinctive correlational procedures (within the strategy of single variations). That is, in R technique one begins on the variables axis and sums over persons: in Q technique one begins on the persons axis and sums over variables (see Figure I.6/1).

One strategy of double variation combines two of Cattell's dimensions -- variables and times -- and sums over the third, persons. This happens to be one of the strategies (otherwise known as the cross-lagged correlation) that sometimes permits causal inference. Another strategy combines concepts and populations (the paradigm of external validation). Another combines populations and times

Figure I.6/1. Cattell's Data Cube (1952).



The Six Correlational Techniques:

R - Correlating variables, replication over persons.

P - Correlating variables, replication over occasions.

Q - Correlating persons, replication over variables.

S - Correlating persons, replication over occasions.

T - Correlating occasions, replication over persons.

O - Correlating occasions, replication over variables.

(the paradigm of direction of influence from one population to another). Etc. (see Figure I.6/2).

As is true of Cattell's six techniques, not all of our ten combinations are equally interesting. Not all are represented in the research literature. However, a systematic review prompts one to wonder why the gaps exist -- because the combinations are really unpromising, or because researchers haven't yet explored their potential for validation and substantive inference?

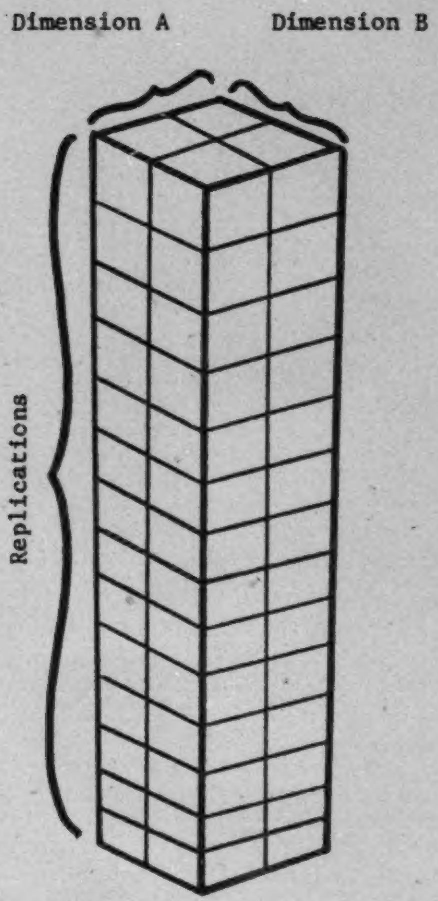
I.7. Convergent versus Discriminant Uses for the Matrix

For each matrix, there should be a convergent application and a discriminant application. In the case of convergence, as in convergent validation, discrimination strengthens the argument by showing that convergence is not merely an artifact of positive relationships that pervade the data (e.g., from an acquiescence set). In the case of discrimination, as in cross-lagged correlation, convergence strengthens the argument by showing stability in matrix cells in which discrimination is not sought (e.g., in cross-lagged correlation, the "test-retest" correlations).

Convergent uses of the matrix are chiefly methodological, ranging from reliability and robustness of technique to validity and generalization across times and populations. Discriminant uses of the matrix are chiefly substantive; usually some inference of effect is the goal.

The particular correlation coefficients showing convergence and discrimination will be discussed in Section III. The point to be made here is that each matrix will sometimes be examined in convergent mode and sometimes in discriminant mode. Thus, although

Figure I.6/2. Data Solid for the Convergent-Discriminant Matrix.



Since there are five classes of variables for dimensions A and B -- concepts, measures, populations, times, and analysis models -- there are ten basic convergent-discriminant matrices: CM (concept by measure), CP, CT, CA, MP, MT, MA, PT, PA, TA. Replication may be over any available dimension, such as populations or persons, times, measures, concepts.

there are only ten matrices in terms of structure, there are twenty matrices in terms of function. And -- to anticipate Section II.1 -- alternative ways of obtaining replicates raise the total number of matrices to eighty or more.

II. TECHNICAL ISSUES

II.1. The Need for a Third, or Replication, Dimension

The decision to study concepts across measures, as in the multitrait-multimethod matrix, is incomplete until we have specified how we shall obtain pairwise observations needed for correlation. That is, what unit of observation permits us to say that X_i and Y_i comprise a correlatable pair while X_i and Y_j do not? Do *i* and *j* refer to different people at the same point in time? Or different times in the same person? Or, as in Q technique, different traits in the same person?

We know that X and Y themselves are defined by the dimensions being compared in the convergent-discriminant matrix. Clearer symbols for the concept-measure matrix would be $C1M1_i$, $C1M2_i$, $C2M1_i$, $C2M2_i$, $C1M1_j$, etc., since each datum refers simultaneously to two dimensions, to a unique combination of concept and measure. The choice of dimension on which *i* differs from *j* remains open, however. This third, or replication, dimension is a key to the versatility of the convergent-discriminant matrix.

Conventionally, in the multitrait-multimethod matrix, concepts and measures are compared with *persons* as replicates. Pairs of observations are correlatable because each pair is obtained from one person. There is nothing wrong with this conventional approach, but note that a multitrait-multimethod matrix could also be computed with timepoints as replicates and with only one person providing all data. A time-replicated multitrait-multimethod matrix has different uses from a person-replicated multitrait-multimethod matrix, as will be discussed in Section III.

Beyond these two obvious variations, multitrait-multimethod matrices can also be computed with other concepts or measures as replicates. This is a form of profile analysis in which other concepts or measures create a profile of measurements for a trait-method combination in lieu of direct observations across persons or times. Replication may also come from matched aggregates of individuals in the population; this approach follows the tradition of ecological correlation, although important qualifications will be discussed. The following section presents both profile analysis and aggregate correlational techniques, because of the special problem of correlating independent data sets (i.e., different populations, etc.). Several convergent-discriminant matrices using these types of replication are presented in Section III.

II.2. Correlating Independent Data Sets

Sometimes we lack the pairwise observations needed for correlation. For example, we may wish to study concept-concept relationships across time in the same population, when different samples of the population provide data on each occasion. Or we may wish to study concept-concept relationships in two populations. In these cases there is no continuity of unit, across times or across populations, to link observations in correlatable pairs.

Another example is the comparison of analysis outcomes. Suppose that we wish to compare agreement among alternative analysis models applied to the same data. With certain exceptions, as when residuals are computed in regression analysis and factor scores in factor analysis, the individuality of data-providing units is lost in analysis. Therefore, differences in analysis outcomes generally cannot be imputed to individual cases for correlational treatment.

At other times correlatable pairs exists, but on another replication dimension besides "persons" or their analog (as in Cattell's R, S and T models). That is, sometimes correlatable pairs are formed from the same concepts measured in different persons or from the same concepts measured on different occasions in the same person (Cattell's Q, O and T models). In these cases correlations are being computed across a profile for a concept, measure, population, time, or analysis model -- a profile provided by other concepts, measures, times, population, and analysis models.

Profile correlation. Psychometricians define profile as a "set of different measures of an individual or group, each of which is expressed in the same unit of measure" (Kerlinger, 1964, p. 614). Although this definition refers only to "measures," we can readily extend it to other potential replication dimensions, provided we consider carefully the implications of other parts of the definition.

The first criterion for a profile is that all constituent scores of the profile must refer to a single (conceptual) unit of analysis. Just as all scores in a personal profile are provided by an organically unified subject, all scores in a population profile must represent the population as a unity. When the measure is not a composite of individuals' scores (e.g., in the case of an index, such as GNP), we must be able to make it applicable to the entire population being profiled. On the other hand, when the measure is a composite of individual measures, we should have some confidence that the composite is representative of the entire unit of analysis.

In sum, twenty concepts measured in a single random sample of a population would constitute a profile of that population. However, one concept measured in twenty non-random aggregates of the same population -- whether or not the twenty groups together exhaust the population -- does not; this situation fails the uniform-unit criterion of profiles. The case in which correlatable pairs are non-random sub-groups that exhaust a population is discussed below as aggregate correlation.

The second criterion implied in our definition of profile is that the constituent scores should all be expressed in the same unit of measure. Two aspects of this criterion must be considered.

First, while all scores in a profile need not be measured in the same metric (e.g., inches, feet, seconds, dollars), all scores should refer to the unit of analysis (e.g., per capita, or percentage of the population). Thus, a conceptually uniform profile of one of the 50 states might include median education level, value that manufactures add to the state in a given year, per capita expenditures on state highways, per capita income, per capita electric energy production, etc. Some scores, (e.g., median education) do not require transformation into per capita terms, but variables like state highway expenditures clearly do.

Second, once scores have conceptual uniformity, they should be made statistically uniform by the assignment of rank-order scores or standard scores. Both rank-order scores and standard scores require a distribution of populations, not just two, in order for the ranks or the means and standard deviations to be meaningful, respectively.

Having considered the criteria for scores constituting a profile, let us now consider some possible approaches to correlation when the replication dimension forms profiles of independent data sets.

A first approach is a matching on equivalent concepts or measures. For example, populations can be matched on indices of social and economic development (e.g., infant mortality, calories of food consumption, kilowatt hours of electricity, nonagricultural workforce, gross national product). The criterion of equal units of measurement is important in the use of such multiconcept or multimeasure indices. It is essential to consider both aspects of the problem, conceptual uniformity and statistical uniformity.

We are begging the question of whether social and economic indices are measures or concepts. If the study involves a global concept such as "level of modernization," the indices may only be multiple measures. Nonetheless, "gross national product" can itself be a concept, operationalized in various ways.

Populations can also be matched on other populations. For example, the similarity of international policy between populations A and B might be measured by correlating the number of diplomatic contacts each has with other nations of the world (assuming that populations A and B are nation-states and have international policies). Similarity could also be measured by correlating the deployment of embassies and consulates in other nations, etc. Again, we are begging the question of whether such variables represent matching on other populations or on concepts. It may be only a matter of conceptual appropriateness in any given case.

Finally, populations can be matched on times, with all data provided by the same measure, repeated over time. An example might be the correlation of "economic optimism" cycles in two or more populations based on annual totals of investment or new construction.

All of these examples use raw data that permit the correlation of independent data sets. It is also possible to correlate independent data sets on the basis of second-order or derivative data, such as correlation coefficients themselves. For example, we can imagine a situation in which two tests have never been used jointly -- hence their direct correlation is unknown -- and yet have been used individually in the company of a large number of other tests which both have in common. For test A, then, we have a pattern of correlation with tests I, J, . . . N. For test B we have an equivalent pattern. The indirect correlation between tests A and B, computed from the pattern of their correlation with the same other tests, should approach the direct correlation as the number of other tests becomes larger and/or more heterogeneous.

Such correlation of patterns or profiles of correlation coefficients, a subcase of canonical correlation, permits comparison of the outcomes of different analysis models. If we wish to know the agreement in outcome between different models applied to the same data, or the same model applied to different data, we can correlate the outcomes themselves. An example of the first condition might be the choice of Pearson r, Spearman rho, eta, and tetrachoric correlation for a data set marred by skewness. It is possible to compute average levels of these coefficients, to see which tend to underestimate or or overestimate association throughout an entire matrix (cf. Paisley and Ruggels, 1965), but a correlation of the correlations reveals the ability of each statistic to agree with others in "tracking" changes in relationship throughout the matrix. One of the CA (concepts by analysis models) matrices presented in Section III shows the difference between "average level" versus "tracking ability" criteria in choosing a correlational statistic.

An example of the second condition, the same model applied to different data, might be the replication of a multiple regression analysis on fresh but comparable data. The solution in this case is simple: correlate the profiles of betas. Agreement in the profiles of betas will show whether predictors have the same relationship to the criterion variable in both data sets.

Aggregate correlation. Another approach to establishing correlatable pairs in independent data sets cannot properly be called profiling in the sense we have discussed above. This is matching populations on groups. It may often be possible to identify sub-groups within independent populations that are equivalent with respect to characteristics such as demographic variables and geographic area of residence. In these cases, however, the first criterion of profiles will not be met, since the groups formed in this manner are not drawn to be representative samples of the population. All measures in a "true" profile must be based on a representative sample of the entire profiled object. To make this distinction salient, we will speak of aggregates when comparable non-random groups of subjects form the correlatable pairs.

Table II.2/1 shows a hypothetical profiling of two populations on a combination of demographic attributes. Depending on a theoretical rationale that lies beyond this statistical example, the correlation between the two columns might show agreement in "distribution of educational opportunity" between the populations.

Correlating data from aggregates has been used frequently in the literature of political science, sociology and economics under the name, "ecological correlation." This label generally refers to "correlations in which groups (rather than individuals) from the population are the statistical objects" (Robinson, 1950). In the literature the groups are typically individuals aggregated according to geographic and/or political boundaries (e.g., states, districts, census tracts). However, it is reasonable to extend the present discussion to include sub-groups delineated in other ways (e.g., sex - race - SES - income). In both cases, the aggregates, taken individually, are not representative (i.e., random) samples of the population; but taken together they exhaust a representative sample of the population.

Sometimes we may want to compute correlations from aggregate figures as an estimate of the correlation we would have gotten from correlatable pairs of individuals in the two data sets. Correlating aggregates to this purpose brings some built-in liabilities which require careful conceptualization and statistical manipulation. Two of these, especially treacherous in the computation of a convergent-discriminant matrix, will be discussed here.

1. Aggregate correlations of variables cannot be presumed to approximate individual correlations. Several studies (Robinson,

Table II.2/1. Hypothetical Example of the Correlation of Profiles.

	Percentage of each group completing 12 years of education	
	Population A	Population B
Male, white, blue collar father, family income below $5000	27 %	23 %
Female, white, blue collar father, family income below $5000	36	25
Male, colored, blue collar father, family income below $5000	18	21
Female, colored, blue collar father, family income below $5000	22	23
Male, white, white collar father, family income below $5000	41	39
Female, white, white collar father, family income below $5000	56	59
Male, colored, white collar father, family income below $5000	28	25
Female, colored, white collar father, family income below $5000	34	37
Male, white, blue collar father, family income above $5000	38	33
Female, white, blue collar father, family income above $5000	48	48
Male, colored, blue collar father, family income above $5000	25	21
Female, colored, blue collar father, family income above $5000	29	28
Male, white, white collar father, family income above $5000	74	81
Female, white, white collar father, family income above $5000	80	78
Male, colored, white collar father, family income above $5000	52	48
Female, colored, white collar father, family income above $5000	55	57

1950; Goodman, 1959) have noted the reason for this: aggregate correlations are based on group totals, or marginal values, of the variables, while individual correlations reflect the smallest components of those totals; and many combinations of components can produce exactly the same marginal figures. In other words, "there are a large number of individual correlations which might correspond to any given ecological correlations, i.e., to any given set of marginal frequencies" (Robinson, 1950, p. 354).

Goodman has shown that, in general, aggregate correlations overestimate individual correlations. His proof is quite detailed and will not be repeated here; interested readers may refer to his paper, "Alternatives to ecological correlation" (American Journal of Sociology, 1959, 44, May, 610-25). The important point is that we cannot safely use aggregate correlations to make inferences about the behavior of individuals.

Fortunately, proposals have been advanced that suggest methods of estimating individual correlations from aggregate figures. The most viable of these for the convergent-discriminant matrix appears to be Goodman's suggestion that, when the assumptions for regression are met by the aggregate data, an analysis of regression between the aggregate variables can be used to make inferences about the population of individuals (Goodman, 1959, pp. 620-621). In other words, it is possible to obtain an unbiased estimate of the individual correlation corresponding to an aggregate correlation between two variables, X and Y, for which the regression equation, $Y = bX + a$, has been solved.

Let us consider two concepts, C1 and C2. In Population 1, both are measured by one method (e.g., self-report); in 2, both are measured by another method (e.g., observer's report). Thus, in Population 1 we have only C1M1 and C2M1, while in Population 2 we have C1M2 and C2M2. Since the M1 appears only in Population 1 and M2 only in Population 2, there are no correlatable pairs of individuals. In order to correlate all concepts by all measures with each other, it is necessary to find comparable groups in the two populations and compute correlations on the basis, for example, of group means. Four regressions are necessary to produce a CDM: both C1M1 and C2M1 must be regressed on both C1M2 and C2M2. For each pair, the slope b and the y-intercept a must be computed.

Then, the individual correlation, r_i, can be estimated by multiplying the estimated slope, $\hat{b}$, by the ratio of the population standard deviation of the X variable to the population standard deviation of the Y variable.* In other words, after the regression

---

*Goodman (1959, pp. 620-21) suggests a method of estimating the population variances when the variances of the aggregates are known:

$$\hat{\sigma}^2 \text{ pop.} = \sum w_i M_i$$

where:

i = any of 1 . . . k aggregates;

$w_i = n_i/N$

$M_i = S^2_i + (\bar{x}_i - \bar{X}_i)^2$

where:

S^2_i = variance of group i

$\bar{x}_i$ = mean of group i on variable j

$\bar{X}_i$ = mean of population on variable j equation has been solved, $\underline{r_i}$ can be estimated in the following way:

$$\hat{r}_i = \hat{b}V$$

where:

$$V = \sqrt{\frac{\sigma^2_X}{\sigma^2_Y}}$$

Goodman insists that the assumptions of linearity and homoscedasticity must be met for the aggregate data in order for this technique to yield an unbiased estimate of $\underline{r_i}$. He suggests plotting the data to check for linearity before a regression is computed. In the case of a "2 x 2" CDM, four plots would be necessary. Unfortunately, the technique is anything but robust. Both Cartwright (1969) and the present writers have found it highly sensitive to failure of data to meet assumptions.

Goodman does not report empirical comparisons between correlations computed from individual scores and the estimates obtained from regression scores. However, at least one sample analysis has shown the two to be too discrepant for interchangeability. Test scores from a sample of 2547 high-school students were correlated with individuals as replicates. Another set of correlations was computed with replication over groups, formed by all possible combinations of the levels of some demographic and aspiration variables (viz., father's educational level and type of high-school curriculum enrolled in). The resulting correlations are shown in Table II.2/2. In every case, the aggregate correlations are, as has been predicted, higher than the individual correlations on the same variables.

Table II.2/2. Individual and Aggregate Correlations among Scores for Five Information Tests

	Literature		Music		Math		Art	
	r_i	r_a	r_i	r_a	r_i	r_a	r_i	r_a
Music	617	792						
Math	556	824	550	781				
Art	548	814	519	862	442	768		
Total	683	914	600	742	637	880	725	771

The Goodman technique was applied to the aggregate correlation ($r_a = .824$) between "literature" and "math" information scores in an effort to approximate the known individual correlation for these variables ($r_i = .556$). These variables were chosen because the graph of the relationship between them was generally linear. The resulting estimate of the individual correlation ($\hat{r}_i$) was .709, which is significantly different from the true individual correlation ($p < .05$). However, it is also significantly different from the aggregate correlation ($p < .05$). So while the Goodman technique did not yield a very close approximation to the real individual correlation, it did produce a significantly lower estimate than the aggregate correlation itself. Perhaps the failure of the technique was due to imperfect linearity in the relationship between the two variables.

We have discussed the case in which both variables were quantitative (i.e., measured on at least an interval scale). Goodman also discusses the cases in which one variable is quantitative and the other dichomotous and in which both are dichotomous. These will not be discussed here, since they are less generally relevant to the CDM than the case of two quantitative variables. Interested readers should refer to Goodman's paper.

Cartwright (1969) reviews other methods, as well as proposing one of his own. However, none of them compute a single best estimate using data that would ordinarily be available from a correlation computed on group data.

2. A second caveat in using aggregate correlations has to do with the relationship between the aggregating principle (e.g., geographic areas, demographic variables) and the criterion variable(s). What is the optimal relationship between the two in each set of scores to be correlated? That is, what is the set of confidence limits within which the degree of association can vary without damaging our ability to approximate individual correlations from aggregate correlations? Goodman (1959) suggests that regression analysis, at least, is inadvisable when the aggregating principle and criterion variables are highly related. The optimal relationship between variables and aggregating principles is clearly of small magnitude (but not zero, in which case criterion variable means within the aggregates would deviate from the criterion variable grand means only by chance, and the criterion variables would seem to be uncorrelated). However, empirical studies are needed to indicate reasonable estimates of the tolerable degree of association between them in computing aggregate correlations.

Let us now summarize this discussion of the correlation of independent data sets:

1. Sometimes dimensions being compared in a convergent-discriminant matrix lack the pairwise observations needed for correlation. This is especially true when we are comparing different populations or the outcomes of different analysis models.
2. The problem is to find a basis for pairing observations across the independent data sets. Two general types of approaches to this problem are recognized:
   a. Populations may be profiled in terms of several variables, each of which is representative of the entire population. Profiled populations may be matched on equivalent concepts or measures, other populations, times, or patterns of correlation (internally, or with external criteria).
   b. A population may be represented by aggregates of individuals that are comparable to aggregates in another population (i.e., the principle on which the aggregates are formed -- geographical boundaries, demographic variables -- is the same in the two populations). That is, populations may be matched on equivalent groups. Aggregate scores are not representative of the entire population in the way that profile scores are. When aggregate figures are correlated in

an effort to estimate individual correlations,
a regression analysis (Goodman, 1959) may yield
an unbiased estimate of the individual correlation.

To a large extent, the versatility of the convergent-discriminant matrix is a function of the variety of replication dimensions that can be used to fill a single (e.g., concepts by measures) matrix. Many of these replication dimensions involve independent data sets, and the *logic of matching* is needed to make correlation possible.

II.3. Choosing a Correlation Coefficient

One analysis model is implicit in every convergent-discriminant matrix: it is necessary to fill the matrix with a statistic that can show both convergence and discrimination. All the uses of a convergent-discriminant matrix depend on the robustness, sensitivity, and freedom from bias of the chosen statistic.

Unless there are metric difficulties -- skewness, kurtosis, bimodality -- Pearson product-moment correlation should probably be a first choice. Spearman rank-order correlation may be more robust when there are metric difficulties, but of course it is insensitive to size-of-interval differences that may be important.

Two commonly used correlational statistics, *eta* and the contingency coefficient, lack a negative range; they cannot be accepted in a convergent-discriminant matrix. Two statistics that require dichotomization of data, tetrachoric and fourfold point correlation (*phi*), may suppress important differences among distributions.

Although other statistics for filling a convergent-discriminant matrix can be proposed and defended, it seems best to prepare the raw or derivative data for Pearson product-moment correlation when possible, falling back on rho or tau if metrics are distorted enough to invalidate r.

II.4. A Terminology for Discussing the Convergent-Discriminant Matrix

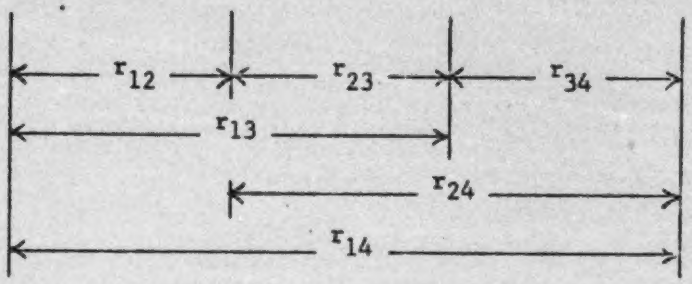
It is the spirit of this paper to preserve the logic and language of the multitrait-multimethod matrix while extending its uses. All terms used to describe the multitrait-multimethod matrix are applicable to at least one convergent-discriminant matrix (concepts by measures). Analogues of these terms are needed for other matrices.

Figure II.4/1 shows a generalized convergent-discriminant matrix, with the two dimensions designated only A and B. The top half of the figure shows the data array from which correlations would be computed. Extending along a third, or replication, dimension, there is an observation (x) for each combination of levels of dimensions A and B, for each replicate. The subscripts of observation x_{123} represent the first level of dimension A, the second level of dimension B, and the third replicate. The correlations between columns (r_{12}, etc.) are keyed to the convergent-discriminant matrix in the bottom half of the figure.

Terminology. Let the convergent-discriminant matrix itself be called a CDM. The major diagonal of the CDM contains reliability values. Since correlations on the major diagonal, when rarely they are available, join the same levels on each dimension, they are referred to as mono-mono values (e.g., monoconcept-monomeasure, monopopulation-monotime, monomeasure-monomodel). Subscripting practice (r_{11}, r_{22}, r_{33}, r_{44}) confirms the reliability function of this diagonal.

Figure II.4/1. General Form of the Convergent-Discriminant Matrix.

		Level 1 of dimension A, level 1 of dimension B	Level 1 of dimension A, level 2 of dimension B	Level 2 of dimension A, level 1 of dimension B	Level 2 of dimension A, level 2 of dimension B
Replication over a third dimension	1	x_{111}	x_{121}	x_{211}	x_{221}
	2	x_{112}	x_{122}	x_{212}	x_{222}
	3	x_{113}	x_{123}	x_{213}	x_{223}
	.	$x_{11.}$	$x_{12.}$	$x_{21.}$	$x_{22.}$
	n	x_{11n}	x_{12n}	x_{21n}	x_{22n}



		A1		A2	
		B1	B2	B1	B2
A1	B1	r_{11}			
	B2	r_{12}	r_{22}		
A2	B1	r_{13}	r_{23}	r_{33}	
	B2	r_{14}	r_{24}	r_{34}	r_{44}

Each CDM contains mono-hetero values in which the level on one dimension remains constant while levels change on the other dimension. A mono-hetero value might relate two measures of the same concept (monoconcept-heteromeasure) or one measure of two concepts (heteroconcept-monomeasure). In another matrix a mono-hetero value might relate two populations at the same time (heteropopulation-monotime) or the same population at two times (monopopulation-heterotime). In Figure II.4/1 the mono-hetero values are r_{12}, r_{13}, r_{24}, and r_{34}.

In Campbell-Fiske terminology, r_{13} and r_{24} would be known as the validity diagonal, assuming that A1 and A2 are methods and B1 and B2 are traits. Whenever we are examining a CDM in convergent mode (see Section I.7), there will be validity or convergence values within the mono-hetero set, but whether they are r_{13} and r_{24} or r_{12} and r_{34} depends on mere mechanics of presenting the matrix. If, in this example, we kept A as the "outside" dimension but identified A with traits and B with methods, then r_{12} and r_{34} would be the validity values, even though they do not form a diagonal. When a matrix represents dimensions such as populations and times, there is no *a priori* validity diagonal in convergent mode. It matters whether convergence is sought across populations, across times, or both.

Finally, each CDM contains hetero-hetero values in which levels and dimensions are completely "out of phase". In Figure II.4/1 r_{14} joins level 1 of both dimensions with level 2 of both dimensions. The two sides of the correlation have no element in common; this can be shown by changing the subscript to $r_{A1B1-A2B2}$. The other hetero-hetero value in Figure II.4/1, r_{23}, joins the A1B2 and A2B1 combinations. Again the two sides of the correlation have no element in common.

Hetero-hetero values, such as heteroconcept-heterotime cross-lagged correlations, are most important in discriminant mode. If well-founded in theory and method, they support inferences of causation and direction of effect.

Now that we have identified three kinds of correlations in the CDM, it will be useful to specify their roles in convergent and discriminant modes:

	Convergent Mode	Discriminant Mode
Mono-Mono	Reliability of observations.	Reliability of observations.
Mono-Hetero	(1) One set of mono-hetero values provides the test of validity (convergence). (2) Other sets of mono-hetero values provide proof of discrimination. Example: monoconcept-heteromeasure values test validity, while heteroconcept-monomeasure values test discrimination.	Baseline of normal change on each dimension, used in assessing significance of hetero-hetero (e.g., cross-lagged) correlations. Example: the monoconcept-heterotime and heteroconcept-monotime correlations should show patterns of stability to contrast with differences between the heteroconcept-heterotime correlations for causal inference.
Hetero-Hetero	Easier test of discrimination than mono-hetero. Having no element in common, the two sides of the correlation should be quite independent.	The causal inference correlations.

III. THE TEN CONVERGENT-DISCRIMINANT MATRICES

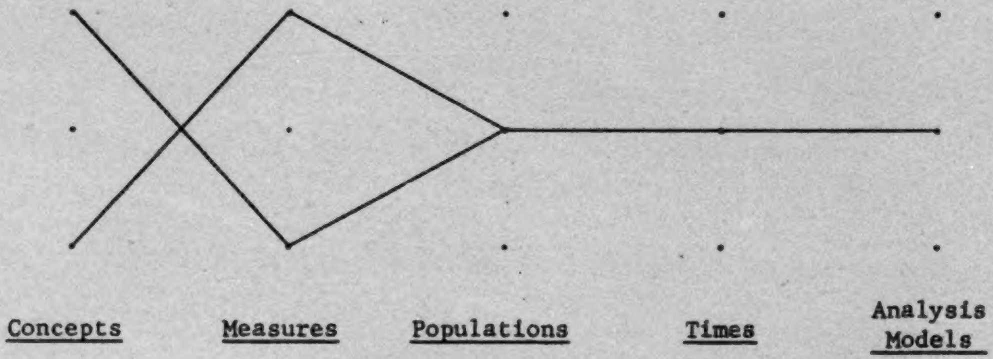
If, as we believe, the idea of the convergent-discriminant matrix is sound, then this illustrative section will always be obsolete. Our hypothetical examples will be superceded by empirical examples, these superseded in turn by better examples. New replication dimensions will be found to make "impossible" comparisons possible. The close affinity of the CDM to nonexperimental methodology in general (_cf_. Campbell and Stanley, 1963) and to unobtrusive measurement in particular (_cf_. Webb _et al_., 1966) should cause the limits of the matrices, especially in discriminant (inferential) mode, to be tested.

As of now, only the CM matrix is well illustrated, thanks to a number of multitrait-multimethod matrices in the literature. Our discussions of the other nine matrices betray uneasiness about operating characteristics that cannot be known until quantities of data have been run. Also, each replication dimension that involves profile or aggregate correlation (see Section II.2) depends on the field's long-term acceptance of assumptions required by those techniques.

Although each of the ten matrices is presented in identical format, with two levels on each dimension, there is of course no limitation on the number of levels that may be compared in a CDM, nor need both dimensions have the same number of levels.

III.1.

The Concepts by Measures (CM) Convergent-Discriminant Matrix

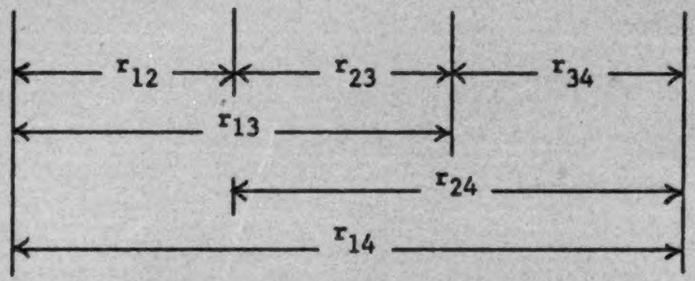


This is the original multitrait-multimethod matrix (see Section I.2). There are two replication dimensions of great utility: populations and times. The first dimension may refer to replication over individual persons, as in most "M^3" examples in the literature, or over larger aggregates of people -- communities, states, nations.

Replication over time implies that just one person is providing data on trait-method (or concept-measure) combinations on several occasions. The special use of a time-replicated CM might be an evaluation of the "tracking" ability of measures. Consider both self-report (M1) and physiological (M2) measures of hunger (C1) and fatigue (C2) at hourly intervals during the waking hours of a week. If the self-report and the physiological measures are able to track hunger but not fatigue over time, then r_{12} will be large and r_{34} small (see Figure III.1/1). If hunger and fatigue become confused in the person's perceptions, while the physiological measures of each remain independent, then r_{13} will be large relative to r_{24}. (Note that the

Figure III.1/1. **Convergent-Discriminant Matrix:** Concepts by Measures (CM)

		Concept 1 Measure 1	Concept 1 Measure 2	Concept 2 Measure 1	Concept 2 Measure 2
Replication over populations (persons) or times, also other concepts and measures	1	x_{111}	x_{121}	x_{211}	x_{221}
	2	x_{112}	x_{122}	x_{212}	x_{222}
	3	x_{113}	x_{123}	x_{213}	x_{223}
	.	$x_{11.}$	$x_{12.}$	$x_{21.}$	$x_{22.}$
	n	x_{11n}	x_{12n}	x_{21n}	x_{22n}



		C1		C2	
		M1	M2	M1	M2
C1	M1	r_{11}			
	M2	r_{12}	r_{22}		
C2	M1	r_{13}	r_{23}	r_{33}	
	M2	r_{14}	r_{24}	r_{34}	r_{44}

lower half of Figure III.1/1 reverses the order of measures and concepts that is customary in the multitrait-multimethod literature. Since convergent-discriminant matrices are interpreted both in convergent and discriminant mode, and since either dimension may be dominant, the CDMs in these figures simply place the first-mentioned dimension in "outside" position. If we were using the CM matrix to validate multiple measures of a concept, r_{12} and r_{34} would be the validity coefficients, not r_{13} and r_{24} on the so-called validity diagonal.)

Campbell and Fiske reviewed the pre-1959 literature in their article. Therefore we present some examples of post-1959 use of the "M^3". These examples typify the ways in which convergent-discriminant logic has been put to use. (For the sake of consistency in Section III, we will call the CM applications, although of course the investigators were using the multitrait-multimethod matrix.)

A recent example from the literature reminds us that the CM matrix provides a stringent test of convergent and discriminant validity -- perhaps a more stringent test than we bargain for. Zuckerman _et_ _al_. (1967) used the multitrait-multimethod matrix to compare five methods for assessing the traits of anxiety, depression, and hostility. The methods were direct judgment (ratings), the Multiple Affect Adjective Check List (MAACL), questionnaire, the Thematic Apperception Test (TAT), and the Holtzman Inkblot Test (HIT). In the case of the HIT, no exact score for depression was available; a corresponding Rorschach docing category was substituted (R*).

The results in Table III.1/1 convict all the measures of excessive common measure variance. There is fair convergence among the ratings, MAACL, and the questionnaire, but without exception there

Table III.1/1. The Correlation of Three Traits or Concepts (Anxiety, Depression, and Hostility) across Five Measures (CM Matrix, Replication over Persons).

		Ratings			MAACL			Questionnaires			TAT			HIT		
		A	D	H	A	D	H	A	D	H	A	D	H	A	R*	H
Ratings	A	(77)														
	D	72	(77)													
	H	-03	-15	(72)												
MAACL	A	64	44	-06	-											
	D	62	49	-13	94	-										
	H	45	25	-05	81	81	-									
Quest'res	A	55	57	14	59	56	27	-								
	D	66	59	-15	68	69	34	84	-							
	H	26	23	40	38	33	25	52	24	-						
TAT	A	06	17	09	14	17	10	09	08	28	(61)					
	D	15	23	09	30	29	27	34	34	14	45	(81)				
	H	-02	02	06	-06	-06	09	-21	-23	-09	41	40	(77)			
HIT	A	-01	04	-09	-05	-05	05	-23	-17	-05	01	32	32	-		
	R*	-24	-29	22	-06	-13	04	-35	-42	06	22	01	19	35	-	
	H	00	-02	-10	-08	-03	-01	-01	-01	10	26	39	24	43	07	-

are heteroconcept-monomeasure correlations higher than the monoconcept-heteromeasure values of the validity diagonal. The MAACL particularly suffers from common measure variance, so much so that concept scores on the MAACL correlate very highly among themselves and correlate about equally well with same and *different* concepts measured by other techniques.

For the two projective tests there is little evidence of either convergence or discrimination among the three concepts. Heteroconcept-monomeasure values for the TAT and HIT are positive and fairly high, certainly higher than the monoconcept-heteromeasure validity values that the two tests share. If we think of the ratings, MAACL, and the questionnaire as coming closest to the central constructs of anxiety, depression, and hostility (a supposition that is fairly strong by the convergence criterion but weak by the discrimination criterion), then it is disturbing to see that the highest validity value joining the TAT or HIT with any of the nonprojective tests is a *minus* .42. In fact, negative "validity" coefficients are the rule rather than the exception for the HIT.

These findings lead Zuckerman *et al*. to conclude: (1) As for methods, the more simple and direct techniques appear to outperform the more complicated, indirect, projective techniques. (2) As for traits, anxiety and depression managed to converge without much discrimination, while hostility was fairly well discriminated from the other traits but failed to converge.

Illustrating the use of the CM matrix in the analysis of a "transition study" (when a new test replaces an older, and presumably inferior, one), Maher, Watt, and Campbell (1960) compared the validity of projective modifications of the Rotter Incomplete Sentence Blank and structured tests of attitude toward "home and parents" and toward "law and justice" in a prison population.

The picture is one of good validity (see Table III.1/2). The diagonal values of .51 and .50 are evidence of convergence of methods. Off-diagonal values of .13 and .12 show concept discrimination via the "easier" heteroconcept-heteromeasure criterion, and the correlations of .05 and -.04 are surprisingly good indicators of concept discrimination via the "harder" heteroconcept-monomeasure criterion.

The authors' interpretation of the matrix seems just right: "In [this] analysis the two types of test have been mutually validating, rather than showing one method to be more valid than the other. The validity coefficients of .51 and .50 reflect equally on both methods. Only in the magnitude of the irrelevant method-variance, in the heterotrait-monomethod [heteroconcept-monomeasure] *r*'s, could the two methods have shown differential validity -- and here no differences appeared. What is needed to compare the relative validity of the two methods is a third method against which they could each be checked. Ratings by guards or cellmates on the two attitudes would have done well for this purpose, for example." (1960, p. 286).

Table III.1/2. The Correlation of Structured and Projective Tests of Attitudes toward "Home and Parents" and "Law and Justice" (CM Matrix, Replication over Persons).

		Structured		Projective	
		H&P	L&J	H&P	L&J
Structured	H&P	--			
	L&J	05	--		
Projective	H&P	51	12	--	
	L&J	13	50	-04	--

In a five-concept, four-measure study, Norman (1963) attempted to validate peer ratings against scores on the Descriptive Adjective Inventory (DAI), the Occupational Preference Inventory (OPI), and the Welsh Figure Preference Test (WFPT). The five concepts, extracted as factors in a series of factor analyses, were the personality traits of Surgency (or Extroversion), Agreeableness, Dependability, Emotional Stability, and Culture. The primary purpose of the study was to reflect upon the "Unimportance of Test Item Content" controversy (*cf*. Berg, 1959). The DAI, OPI, and WFPT were considered to be highly item-relevant, moderately item-relevant, and item-irrelevant measures of the five traits.

Because of space limitations, only the first three traits are shown across four measures in Table III.1/3 (the entire matrix can be found in Norman, 1963). This submatrix represents overall findings: when

Table III.1/3. The Correlation of Three Personality Traits and Four Measures, Submatrix from Original Study (CM Matrix, Replication over Persons).

		Surgency				Agreeableness				Dependability			
		Rat.	DAI	OPI	WFPT	Rat.	DAI	OPI	WFPT	Rat.	DAI	OPI	WFPT
Surgency	Rat.	-											
	DAI	40	-										
	OPI	03	23	-									
	WFPT	00	03	03	-								
Agreeableness	Rat.	08	-23	06	07	-							
	DAI	-21	-31	-02	11	25	-						
	OPI	-04	01	-17	-16	07	07	-					
	WFPT	03	04	-03	75	08	09	07	-				
Dependability	Rat.	-27	-30	-01	10	48	22	09	08	-			
	DAI	-33	-46	-06	09	06	33	-11	-01	40			
	OPI	07	-14	-32	-02	04	05	-25	-07	-08	10	-	
	WFPT	-01	-02	-05	-22	-14	-07	11	-05	-22	-06	-12	-

peer ratings are taken as the criterion for factor structure, only the DAI, with highly relevant item content, shows any convergent-discriminant validity at all. Both the OPI and WFPT show near-zero validities (the underlined monoconcept-heteromeasure coefficients).

Norman concludes, "Taken together, the results appear to argue rather forcibly against Berg's contention that test item content is unimportant for the construction of valid empirical keys." (1963, p. 172).

Borgatta has used the CM in analyzing personality assessment through self-ratings, self-rankings, and peer-rankings. An early study (1960) in this series examined, for 99 college males, the correlations among self-rating and self- and peer-ranking in both 3-person and 5-person discussion groups, across factored traits labeled assertiveness, sociability, and emotionality. The results showed, in general, a two-fold stability pattern. One is across the three traits for self-ratings and self- and peer-rankings in both the 3- and 5-person groups. The other is across the five ratings for the three traits.

A replication of this study (1964) expanded the three traits to five (assertiveness, likeability, emotionality, intelligence, and responsibility) and obtained the self- and peer-rankings in intimate and stratified group settings. In other words, for 313 female and 144 male subjects, there were five methods of obtaining scores and there were five measured traits.

Table III.1/4 presents an extract of Borgatta's very large matrix: five measures of three traits for the males only. We find in it:

1. The validity diagonals are significant, with few exceptions.

Table III.1/8.

Multitrait-Multimethod Matrix for Five Measures of Three Traits

	Intimate Group						Stratified Group								
	Peer-rankings			Self-rankings			Peer-rankings			Self-rankings			Self-ratings		
	A	L	I	A	L	I	A	L	I	A	L	I	A	L	I
Intimate Group, Peer-rankings															
Assertiveness	-														
Likeability	-02	-													
Intelligence	19	28	-												
Intimate Group, Self-rankings															
Assertiveness	65	-14	07	-											
Likeability	-06	43	-07	02	-										
Intelligence	29	-09	46	23	02	-									
Stratified Group, Peer-rankings															
Assertiveness	59	-13	01	43	-01	18	-								
Likeability	-06	36	21	02	08	-10	03	-							
Intelligence	16	10	47	08	-05	28	39	33	-						
Stratified Group, Self-rankings															
Assertiveness	46	-12	-03	57	-03	22	61	01	28	-					
Likeability	-09	30	-06	-07	50	-09	-02	19	-11	-02	-				
Intelligence	13	-11	30	06	-05	53	25	-05	47	30	01	-			
Self-ratings															
Assertiveness	47	-22	-02	63	-06	13	43	-12	12	58	-09	03	-		
Likeability	07	17	-14	16	43	-07	11	-01	-19	11	46	-18	33	-	
Intelligence	21	03	29	20	02	58	13	-12	24	25	-04	39	39	36	-

2. Each validity value is higher than any value lying in its vectors in the heteromeasure block, again with few exceptions.
3. There is a strong trend for a trait to correlate more highly with itself across methods than with other traits in the same measure block.

Borgatta's conclusion alludes to the complete matrix: "Peer-rankings appear to predict other peer-rankings; self-rankings tend to predict self-rankings; self-ratings tend to predict self-rankings; and self-rankings and self-ratings appear to predict peer-rankings." Specifically, averaged validity diagonals across the male and female data show that intimate peer-rankings correlate .47 with stratified peer-rankings; intimate self-rankings correlate .57 with stratified self-rankings; self-ratings correlate .59 and .53 with self-rankings and .32 and .25 with peer-rankings; and self-rankings correlate .45, .31, .26, and .45 with peer-rankings (the higher two being within a setting -- intimate only or stratified only). Building upon these findings, Borgatta (1964) proposed a Behavioral Self-Rating Form (BSR), a short test useful when time must be conserved.

Utilizing a CM matrix to examine correlations among behavioral disorder scores given by several types of judges to children living in residential treatment centers, Fanshel, Hylton, and Borgatta (1963) reached these conclusions:

1. The validity diagonal joining social workers with non-professionals contains values that are significant and the highest in their vectors. That is, there is convergent validation of behavioral disorder scores

given to the children by these two groups of judges.

2. The validity values joining social workers with psychiatrists, psychologists, and teachers are all positive, but only 30 per cent are highest in their vectors.
3. Although limited by the number of cases, there is evidence of convergent validity between social workers and administrators.
4. Significant heteroconcept-heteromeasure correlations are found, but these are explained by true non-zero correlations among the disorder scores.

In this study, careful examination of the three matrices shows both that a baseline of validity has been established and that differential scoring across judges is present, due perhaps to education and other background factors. Clearly the CM matrix here opens the way to a systematic study of judges' biases in institutional settings.

Wallace and Sechrest (1963) employed the CM matrix to investigate one type of scoring (based on the frequency hypothesis) for four traits (hostility, somatic concern, achievement concern, and religious concern) and five methods (Rorschach, TAT, sentence completion, self-description, and peer reputation). In this study we see the question not only of convergent-discriminant validation but also of choice of best method for measuring a particular trait. Across the four traits, peer reputation and self-description appear to meet validity criteria. An encouraging check is that both methods correlate significantly with behavioral measures. Among the projective measures, sentence completion seems to have acceptable discriminant validity. The TAT and Rorschach do not, however.

Turning to the second issue, it seems that validity values for particular traits differ across methods. As measures of somatic concern, self-description and peer reputation are correlated .59, while self-description and the Rorschach are correlated -.10. Traits high in social desirability, such as religious concern, are perhaps best measured by an instrument like the TAT, while socially undesirable traits, such as hostility, are perhaps best measured by the Rorschach.

In conclusion, Wallace and Sechrest point to method differences and suggest an investigation of the hypothesis "that projective measures tap an aspect of behavioral disposition of which the actor may be at least partially aware but which is either so transformed in overt response or is so subtle in its manifestation that its impact is nil."

Occasionally we find in the literature CM matrices that are not labeled or discussed as such. One of these (from Moran, 1966) is a correlation matrix for two word-association lists which were balanced according to the type of "idiodynamic" response set (functional, concept-referent, or contrast) elicited from a control group.

Each list was read to 482 college students at five-second intervals. Table III.1/5 shows correlations for the three sets and the two lists.

Table III.1/5. CM Matrix for Word Association Response Sets.

		List A			List B		
		F	*CR*	*C*	*F*	*CR*	*C*
List A	*F*	-					
	CR	-08	-				
	C	-33	-07	-			
List B	*F*	39	-06	-24	-		
	CR	02	53	-05	-24	-	
	C	-22	-02	73	-34	-08	-

Drawing upon the matrix and the text of Moran's article, we might conclude:

1. The monoconcept-monomeasure reliabilities are reported to range from .56 to .84 -- in the adequate range.
2. The three traits display the same pattern of correlation in both lists.
3. The several significant values in the heteroconcept-heteromeasure triangles indicate a lack of trait independence. This is not surprising, since the response sets compete for dominance; if one is high, the others must be low (as the negative correlations attest).
4. The monoconcept-heteromeasure validity diagonal values are significantly high, higher than any other correlations lying in each value's vector. Therefore we can safely infer some stability of response set, per person, across word lists.

The last example of CM application is our own, with international development data drawn from the Handbook of Political and Social Indicators (Russett et al., 1964). The two concepts were "newspaper diffusion" and "radio diffusion" in 46 to 91 countries. It was expected that the print medium and the broadcast medium would have somewhat different diffusion patterns, although both are strongly related to national levels of economic development.

It would be possible to correlate directly the newspaper and radio data for each country. However, both as an exercise in profile analysis and as a test of the effects of skewness and missing data, each data set was prepared for analysis two ways ("complete" versus "estimated"), and profiles were computed for each concept across 18 predictors, using two analysis models (Pearson r and Spearman rho).

Because of the spottiness of international development data, complete records were available for only 46 nations. From the 47th to the 91st nation, it was necessary to estimate the value of one or more of the predictors from regional means, values in neighboring countries, etc. Newspaper and radio values were never estimated; nations lacking those variables were excluded from the analysis. The first 46 nations comprise the complete data set for this analysis. All 91 nations comprise the estimated data set.

We consider the difference between complete data and estimated data to be a measure variation, belonging to the class of "all operations required to fill a data matrix" (Section I.4, above).

Profiles were created for each concept, in each data set, by computing Pearson r and Spearman rho between each of the 18 predictors and the concept. The difference between the analysis models in this case is that of correlating highly skewed raw data versus rankings of those data.

Table III.1/6 shows the profile data from which CM matrix correlations were computed. Except that it contains three dimensions of variation instead of two, this table corresponds to the generic data array in the top half of Figure III.1/1. The names of the 18 predictors are omitted, because most require extended definition, but they include GNP/capita, proportion of the population that is literate, proportion of the population in the non-agricultural workforce, etc.

The profiles tell us that concept variance is almost overwhelmed by measure variance and model variance. Sometimes, as in the case of the 6th replicate, profile differences are accounted for by the model factor alone. In the case of the 15th replicate, the model factor has an opposite effect on each concept. In the case of the 1st replicate, the model factor is apparent in only one concept. The 18th replicate comes as close as any to showing a measure factor -- regular differences between the estimated and complete profiles.

Table III.1/7 contains two CM matrices, one for each analysis model. They are directed to the measurement question, "To what extent is the profile of each concept upset by the procedure of estimating a large number of missing data points, relative to profiles obtained from complete data?" The answer seems to be that, although the concepts are not as distinctive as they should be for a strong test of the extimation effect, agreement between the 46-case complete data set and the 91-case estimated data set is high for each concept (the monoconcept-heteromeasure or validity correlations) when a Pearson $\underline{r}$ analysis model is used.

Table III.1/6. Profiles of Correlations Relating "Newspaper Diffurion" and "Radio Diffusion" to 18 Predictors.

		Newspaper				Radio			
		Estimated		Complete		Estimated		Complete	
		Unranked	Ranked	Unranked	Ranked	Unranked	Ranked	Unranked	Ranked
Replicates	1.	81	84	88	90	65	82	61	86
	2.	05	10	21	23	-01	15	05	26
	3.	00	-06	-06	-06	12	-06	09	-03
	4.	06	-29	10	-02	27	-23	34	01
	5.	12	26	17	-05	00	18	-03	-07
	6.	-36	-87	-38	-87	-35	-86	-34	-86
	7.	65	63	64	67	55	65	43	65
	8.	75	87	82	86	70	83	69	81
	9.	78	93	77	90	91	91	93	94
	10.	64	76	73	79	55	74	58	75
	11.	79	92	80	92	70	87	64	86
	12.	-03	15	-30	-15	10	23	-12	-03
	13.	09	22	05	05	-06	12	-12	00
	14.	-27	-52	-56	-59	-27	-60	-39	-58
	15.	56	73	51	68	74	70	76	69
	16.	61	83	53	72	50	76	43	69
	17.	-79	-86	-82	-87	-73	-81	-72	-86
	18.	02	12	15	25	12	19	34	29

Table III.1/7. The Correlation of "Newspaper Diffusion" and "Radio Diffusion" with Estimated and Complete Data (CM Matrix, Profile Replication).

		Estimated		Complete	
		Newsp.	Radio	Newsp.	Radio
Pearson r analysis model:					
Estimated	Newsp.	-			
	Radio	967*	-		
Complete	Newsp.	976	934	-	
	Radio	932	979	937	-
Spearman rho analysis model:					
Estimated	Newsp.	-			
	Radio	996	-		
Complete	Newsp.	973	978	-	
	Radio	970	980	998	-

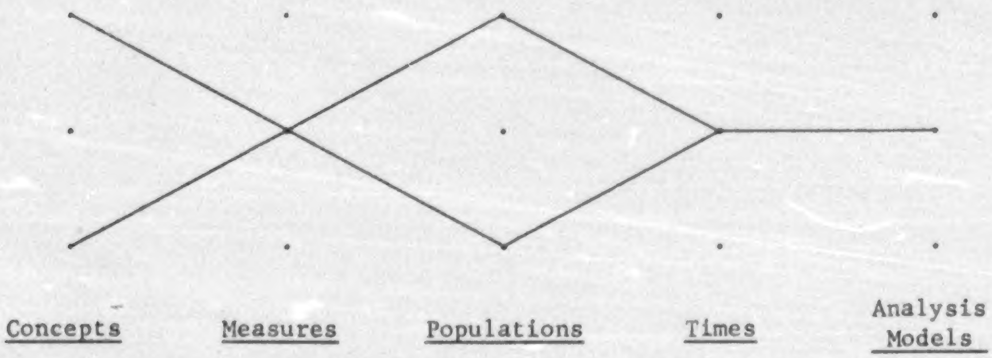
*Correlation to three decimal places, decimal point omitted.

When Spearman rho is used, the heteroconcept-monomeasure correlations are higher than the validity diagonal. Since we can compare the performance of r and rho directly, it is clear that rho is not sensitive to differences between the profiles of the two concepts within the complete and within the estimated data sets. Pearson r may be picking up a size-of-interval difference (e.g., although newspaper and radio both have greatest diffusion in countries with high literacy, newspaper depends more on literacy than does radio and closely follows the level of literacy in each country, whereas radio diffusion merely follows a rank ordering of countries on literacy, differences in literacy proportion between ranks being very disparate).

Further discussion of the difference between analysis models appears in the CA and MA sections. A general conclusion for this CM comparison is that estimating missing data points does not distort concept profiles greatly, at least not to the extent that two concepts within a complete data set have more similar profiles than the same concept in both a complete and an estimated data set. This conclusion holds only for the "standard", or Pearson r, analysis model, however.

III.2.

The Concepts by Populations (CP) Convergent-Discriminant Matrix



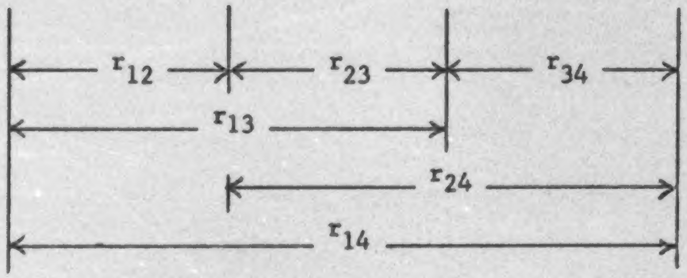
In convergent mode, the CP matrix is unique in providing a test of external validity -- the generality of a concept or relationship across populations.

We present two empirical external validity tests using the CP matrix. Since pairwise observations are not available in raw data across independent populations, both examples are based on correlated profiles of derivative data -- profiles of partial correlation coefficients.

First, from a study of scientific information flow by Parker, Lingwood, and Paisley (1968), a comparison of the stability across populations of two behavioral patterns: the tendency to seek scientific information from interpersonal (other researchers) versus impersonal (journals, books, etc.) sources. In a sample of 786 social scientists, the interpersonal and impersonal "styles" of information seeking had quite different profiles of demographic correlates. Two independent samples of 393 each were created by random division of the total sample, and profiles for each pattern in each sample were derived from multiple regression analyses in which the traits were criterion variables against twelve demographic and other predictors. The profiles consist of twelve 11th-order partial correlation coefficients per trait. Table III.2/1 shows the results when these profiles are correlated within and across populations.

Figure III.2/1. Convergent-Discriminant Matrix: Concepts by Populations (CP)

		Concept 1 Population1	Concept 1 Population 2	Concept 2 Population 1	Concept 2 Population 2
Replication over measures or times, also other concepts and populations	1	x_{111}	x_{121}	x_{211}	x_{221}
	2	x_{112}	x_{122}	x_{212}	x_{222}
	3	x_{113}	x_{123}	x_{213}	x_{223}
	.	$x_{11.}$	$x_{12.}$	$x_{21.}$	$x_{22.}$
	n	x_{11n}	x_{12n}	x_{21n}	x_{22n}



		C1		C2	
		P1	P2	P1	P2
C1	P1	r_{11}			
	P2	r_{12}	r_{22}		
C2	P1	r_{13}	r_{23}	r_{33}	
	P2	r_{14}	r_{24}	r_{34}	r_{44}

Table III.2/1. The Correlation of Two Information-Seeking "Styles" across Populations (CP Matrix, Profile Replication).

		Population 1		Population 2	
		Interpersonal	Impersonal	Interpersonal	Impersonal
Pop. 1	Interpersonal	-			
	Impersonal	20	-		
Pop. 2	Interpersonal	94	27	-	
	Impersonal	15	98	18	-

There is impressive external validity (actually, stability of predictor profiles across populations) in each of the two traits. The very high monoconcept-heteropopulation correlations contrast with the heteroconcept-monopopulation and heteroconcept-heteropopulation correlations, all of which are low.

Much the same results would have been obtained, we think, through the more traditional but cumbersome approach of using each set of twelve betas in multiple regression equations with the other criterion variable and the other population, then computing agreement from the amount of multiple correlation shrinkage.

While the independent random half-sample is a fair test of external validity, a stronger test is obtained from populations that are not only independent but also geographically separate. The next example, a re-analysis of data reported by Verba _et al_. (1967), divides the United States into Eastern and Western halves, for external validation of predictor profiles of "escalation" and "de-escalation" sentiment concerning the Viet Nam war.

A survey of public knowledge about, and attitude toward, the Viet Nam war provided scales of escalation (ESC) and de-escalation (DE-ESC) sentiment. ESC and DE-ESC were each composed of five items, with no overlapping content. Each met Guttman reproducibility criteria. Although apparently opposite in meaning, the scales have only a moderate negative correlation. An attempt to combine them in a single Guttman scale showed that they tapped different dimensions of attitude.

ESC and DE-ESC are interesting concepts for CDM study because we expect a certain relationship between them, across populations, analysis models, or whatever. As in the previous example, correlations were computed across profiles of the predictors of each concept. The 24 predictors included personal attributes of respondents as well as scales of information level and attitude toward Communism and the costs of the war. After the 1495 respondents were divided into East and West subsamples, product-moment correlations were computed within and between subsamples on profiles derived from four analysis models: _r_, _eta_, stepwise multiple regression, and simultaneous multiple regression. That is, one set of profiles consisted of Pearson _r_'s between each of the predictors and ESC and DE-ESC. Another set of profiles consisted of _etas_ between the predictors and ESC and DE-ESC. Two other sets consisted of partial correlation coefficients from stepwise and simultaneous multiple regressions, respectively. The stepwise partials were those of each predictor at the point at which it was accepted into the equation. The simultaneous partials were simply 23rd-order partials.

The data solid for this analysis comprised two concepts by two populations by four analysis models, with 24-predictor profiles of correlation coefficients serving as the replication dimension. Correlations between profiles filled CP, CA, and PA matrices.

We present here only the CP matrix derived from profiles of 23rd-order partial correlations. Other matrices based on these data appear in Sections III.4 and III.9.

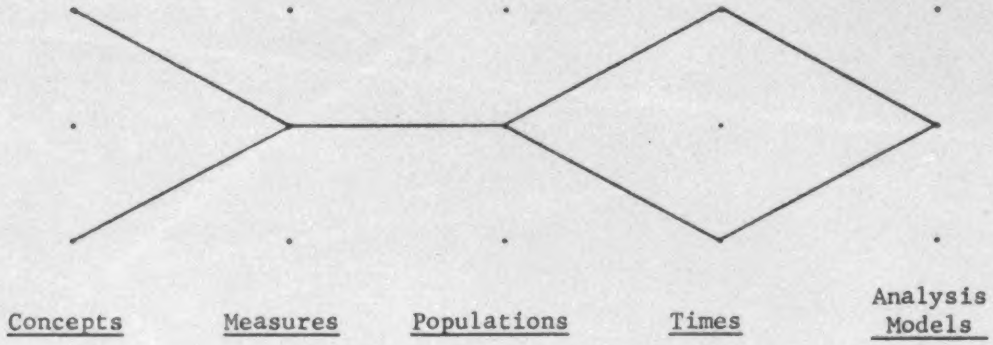
Table III.2/2. The Correlation of Viet Nam Escalation and De-escalation Sentiment across Populations (CP Matrix, Profile Replication).

		Eastern U.S.		Western U.S.	
		ESC	DE-ESC	ESC	DE-ESC
Eastern U.S.	ESC	-			
	DE-ESC	-12	-		
Western U.S.	ESC	62	-18	-	
	DE-ESC	-17	67	-11	-

In Table III.2/2, the two convergence correlations (monoconcept-heteropopulation) and the four discrimination correlations show that ESC and DE-ESC each has a stable set of correlates across populations and that the correlates of one are not the correlates of the other. Stated more concretely: the subgroups that are high in escalation sentiment in one population prove to be the same in the other population (to the extent of the r of .62); subgroups high in de-escalation sentiment in one population prove to be the same in the other population (r of .67); and the pattern of de-escalation subgroups is not merely the converse of the escalation pattern.

III.3.

The Concepts by Times (CT) Convergent-Discriminant Matrix



The time dimension is the most neglected resource in nonexperimental social research. Whereas the experimental researcher may manipulate time along with other variables, most nonexperimental studies have been designed for static, once-only data collection. Since causal inference depends on the time dimension because of its antecedence criterion, nonexperimental research usually -- and unnecessarily -- lacks inferential power.

Just as the CM and CP matrices have unique utility as tests of construct and external validity, the CT and PT matrices are unique in supporting substantive inferences. If theory and method are handled carefully, spurious relationships eliminated, alternative explanations rulled out, etc., then the CT matrix supports the inference that one phenomenon is a cause (not necessarily the cause) of another. Similarly, and with similar qualifications concerning theory and method, the PT matrix supports the inference that one population (perhaps a single person) is changing over time toward the model or profile of attributes earlier exhibited by another population (or person).

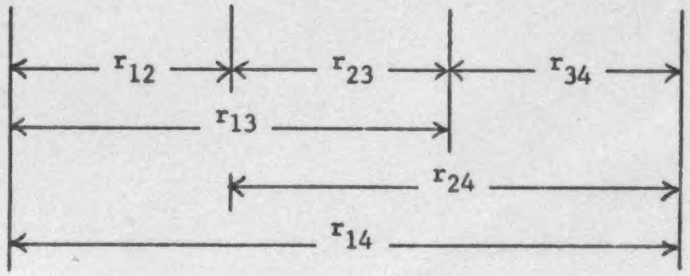
The CT matrix is a more general paradigm for the cross-lagged correlation discussed by Campbell and Stanley (1963) inter alia. The literature of cross-lagged causal analysis is reviewed in Section IV. The following hypothetical example illustrates use of the CT matrix in discriminant mode:

Achievement motivation (C1) and grade-point average (C2) are measured in a population of students at an interval of several years (T1 and T2). The monoconcept-heterotime correlations r_{12} and r_{34} show stability of achievement motivation and grade-point average, individually, over time. The heteroconcept-monotime correlations r_{13} and r_{24} show the amount of association between achievement motivation and grade-point average within each time period. The four reliabilities (r_{11}, r_{22}, r_{33}, r_{44}) show whether any concept-time combination is handicapped in the analysis by excessive error variance.

The heteroconcept-heterotime r_{14} and r_{23} are the cross-lagged correlations. If r_{14} (achievement motivation-T1 by grade-point average-T2) is significantly larger than r_{23} (grade-point average-T1 by achievement

Figure III.3/1. Convergent-Discriminant Matrix: Concepts by Times (CT)

		Concept 1 Time 1	Concept 1 Time 2	Concept 2 Time 1	Concept 2 Time 2
Replication over measures or populations, also other concepts and times	1	x_{111}	x_{121}	x_{211}	x_{221}
	2	x_{112}	x_{122}	x_{212}	x_{222}
	3	x_{113}	x_{123}	x_{213}	x_{223}
	.	$x_{11.}$	$x_{12.}$	$x_{21.}$	$x_{22.}$
	n	x_{11n}	x_{12n}	x_{21n}	x_{22n}



		C1		C2	
		T1	T2	T1	T2
C1	T1	r_{11}			
	T2	r_{12}	r_{22}		
C2	T1	r_{13}	r_{23}	r_{33}	
	T2	r_{14}	r_{24}	r_{34}	r_{44}

motivation-T2), then we begin to infer that achievement motivation leads to successful academic effort. If r_{23} is significantly larger than r_{14}, then we begin to infer that "success breeds ambition" -- successful academic effort leads to achievement motivation.

Causal inference begins with the comparison of r_{14} and r_{23}, but defense of the inference is an open-ended challenge. Is the larger of the two cross-lagged correlations, either r_{14} or r_{23}, adequately large relative to r_{12}, r_{34}, r_{13}, and r_{24}? If, for example, achievement motivation leads to successful academic effort, should grade-point average be more strongly related to achievement motivation at T1 than at T2 (i.e., how large should r_{14} be relative to r_{24})? Beyond the two concepts, is there a prior phenomenon that accounts for both achievement motivation and grade-point average, just as the onset of puberty accounts for both the appearance of acne and an interest in the opposite sex (between which the cross-lagged correlation may be quite high)?

Instead of presenting empirical cross-lagged correlations from the literature (which the reader can easily reach through references in Section IV), we report two original analyses by Collins. The first, shown in Table III.3/1, deals with "production of knowledge" as a function of federal awards to universities. Taking the yearly number of Ph.D. degrees as an index of knowledge production, the question is: do prior federal awards lead to a later Ph.D. increase, or does a prior record of Ph.D. production lead to an increase in federal awards? The data derive from NSF reports of federal awards, by agency, to each university, and variables were added from the Statistical Abstract of the United States so that some effects of the local setting (e.g., SES, industrialization, state expenditures for education) could be controlled. The years 1963 and 1966 were chosen on the basis of availability alone.

Table III.3/1 shows a problem that turns up in CT matrices when variables are reliably measured: correlation of a variable with itself over time (the autocorrelation) can be quite high. For example, the 1963 and 1966 Ph.D. production levels correlate .90, leaving these two variables only 19 per cent independent variance $(1.00 - .90^2)$ for differential effect. Total federal awards are correlated from 1963 to 1966 at exactly the same level, and DOD awards are autocorrelated .94 (it is interesting to note that autocorrelations show the stability -- or inflexibility, however you prefer to characterize it -- of a federal agency's award policies). The lowest autocorrelation is NSF's: .41.

The substantive question is answered to some extent in the heterotime block in the lower left section of the table. First taking only total federal awards, we find that 1963 awards correlate more highly with 1966 Ph.D. level (.76) than the converse (.65). The record by individual agency is less clear, ranging from no difference in the case of DOD (which must be expected because of the strength of the DOD autocorrelation) to a .35 difference in the case of NSF, favoring the award ⟶ Ph.D. production hypothesis.

The second Collins example deals with Statistical Abstract data on the 48 continental states in 1950 and 1960. The key to Table III.3/2 shows the five variables chosen to represent different sectors of development. The general research question is: Which of these development indices "lead," hence may have some causal role, and which "lag?"

Table III.3/1. Knowledge Production in the United States, 1963 and 1966 as a Function of Federal Expenditures, 1963 and 1966*

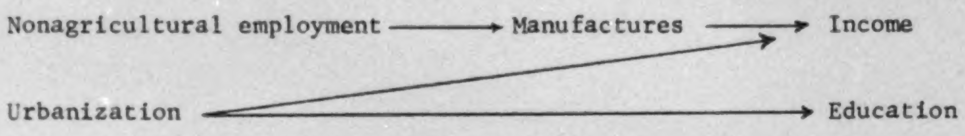
		1963					1966				
		Total	DOD	HEW	NSF	Ph.D.	Total	DOD	HEW	NSF	Ph.D.
	Total										
	DOD	88									
1963	HEW	84	70								
	NSF	84	69	68							
	Ph.D.	76	62	63	77						
	Total	90	69	84	81	65					
	DOD	88	94	69	76	62	81				
1966	HEW	79	50	87	63	54	91	58			
	NSF	54	45	39	41	50	63	60	41		
	Ph.D.	76	61	61	85	90	78	69	62	62	

*Figures represent partial correlations of the major variables, controlling for median education, per capita income, average employment in nonagricultural industries, and state expenditures for education.

Table III.3/2 shows that the monotime correlations (upper left and lower right triangles) have similar patterns and that the monoconcept-heterotime diagonal has the high values of relatively stable and well-measured phenomena. Using the language of correlational causal inference, we say that synchronous correlations exhibit similar patterns across time and that autocorrelations are high.

As for "leading" and "lagging," we find that two variables seem to lead but not to follow. Urbanization and nonagricultural employment are moderately correlated variables that have no differential lagged relationship with each other. However, urbanization slightly "leads" manufactures (.41/.37), nonagricultural employment strongly "leads" manufactures (.76/.49), and urbanization "leads" both income (.75/.58) and education (.55/.43). Nonagricultural employment has no differential lagged relationship with income, but the derivative variable of manufacturing value "leads" income (.39/.28).

Moving from correlations to a directional model with complete disregard for the problems of interpretation that Section IV will discuss, we might propose the following:



Of course, with an analysis this superficial, such a model has no purpose other than to provoke a test.

Table III.3/2. The Correlation of Development Variables and Significant Social Variables in the 48 States, 1950 and 1960 (CT Matrix, Replication over States)

		1950					1960				
		M	U	I	NAE	E	M	U	I	NAE	E
1950	M	-									
	U	38	-								
	I	37	68	-							
	NAE	84	32	24	-						
	E	-08	47	62	-15	-					
1960	M	82	41	28	76	-03	-				
	U	37	85	58	32	43	39	-			
	I	39	75	95	27	63	30	66	-		
	NAE	49	34	29	46	15	51	26	34	-	
	E	-05	55	67	-13	94	-03	45	67	11	-

M = Value added to each state by manufactures

U = Index of population living in urbanized areas

I = Per capita income

NAE = Nonagricultural employment

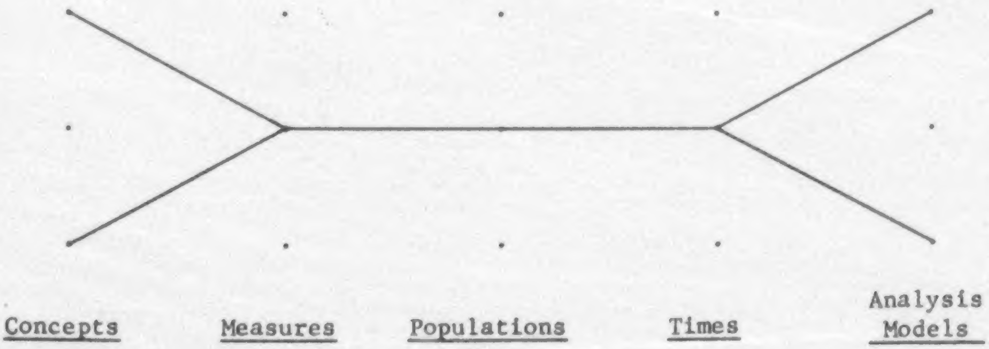
E = Median education

Since CT is perhaps the most promising and elegant of the ten convergent-discriminant matrices, it is important to note that *no CDM has an inherent logic of interpretation*. Different philosophies of science will lead to different interpretations of the most clear-cut CDM patterns.

Turning from discriminant to *convergent* mode, the CT matrix has a minor but interesting role in testing the stability of concepts or relationships over time. CT is serving in convergent mode whenever we are more interested in the succession of heteroconcept-monotime or monoconcept-heterotime correlations than in the cross-lagged heteroconcept-heterotime correlations. If, for example, we hypothesized that artistic creativity increased during periods of personal or social stress and decreased during tranquil periods, we would want to inspect a succession of heteroconcept-monotime correlations -- that is, artistic creativity and stress level within T1, T2, T3, etc. If the two concepts "track" each other, despite changes in both over time as reflected in the monoconcept-heterotime and heteroconcept-heterotime correlations, then our hypothesis is supported more strongly than it would be in a simple time-series analysis lacking the other CDM correlations.

III.4.

The Concepts by Analysis Models (CA) Convergent-Discriminant Matrix



Analysis models, like measures, are researchers' inventions, in contrast to the three "real" (i.e., substantive) dimensions of concepts, populations, and times. *Natural phenomena* could not become *data* without measures, and *data* could not become *findings* without analysis models. The questions to be asked of analysis models, like measures, is, "Are they getting in the way of true trends, failing to track real differences, turning up spurious leads?" The CA, MA, PA, and TA matrices reflect on the performance on analysis models permitting us to test particular models applied to particular concepts, measures, populations, and times.

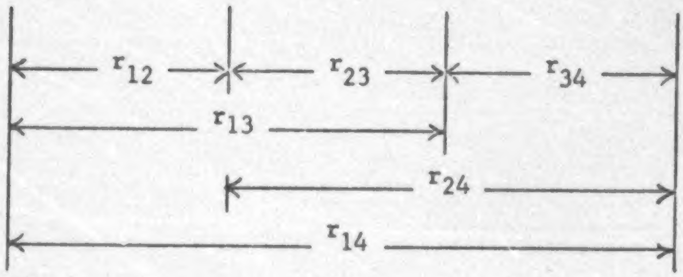
The international development data discussed in Section III.1 provides our first example of a CA matrix. Without repeating the methodology in detail, it will be remembered that two concepts, "newspaper diffusion" and "radio diffusion", were studied with two measurement approaches, a complete data set (46 nations) and an estimated data set (91 nations), using two analysis models, Pearson *r* and Spearman *rho*. Table III.4/1 presents correlations of concepts by analysis models, averaged in each case over two measurement approaches.

Table III.4/1. The Correlation of "Newspaper Diffusion" and "Radio Diffusion" across Analysis Models (CA Matrix, Profile Replication).

		Pearson *r*		Spearman *rho*	
		Newsp.	*Radio*	*Newsp.*	*Radio*
Pearson *r*	*Newsp.*	-			
	Radio	952	-		
Spearman *rho*	*Newsp.*	966	931	-	
	Radio	958	937	997	-

Figure III.4/1. Convergent-Discriminant Matrix: Concepts by Analysis Models (CA)

		Concept 1 Analysis 1	Concept 1 Analysis 2	Concept 2 Analysis 1	Concept 2 Analysis 2
Replication over measures, populations, or times, also other concepts	1	x_{111}	x_{121}	x_{211}	x_{221}
	2	x_{112}	x_{122}	x_{212}	x_{222}
	3	x_{113}	x_{123}	x_{213}	x_{223}
	.	$x_{11.}$	$x_{12.}$	$x_{21.}$	$x_{22.}$
	n	x_{11n}	x_{12n}	x_{21n}	x_{22n}



		C1		C2	
		A1	A2	A1	A2
C1	A1	r_{11}			
	A2	r_{12}	r_{22}		
C2	A1	r_{13}	r_{23}	r_{33}	
	A2	r_{14}	r_{24}	r_{34}	r_{44}

In this table the insensitivity of Spearman rho to differences in concept profiles is as clear as in Table III.1/7. Furthermore, there is poor agreement between two analysis models applied to each concept (monoconcept-heteromodel correlations), relative to comparison values in the matrix.

These negative findings, to the extent that they point to Spearman rho as an insensitive analysis model, contrast with the conclusion of Paisley and Ruggels (1965), who tested rho on similar data together with Pearson r, eta, tetrachoric correlation, and a transformation procedure culminating in Pearson r. Paisley and Ruggels found that Spearman rho was second only to the "transformation + r" procedure in coming out middlemost in the distribution of five correlations for each pair of variables. This tendency neither to overestimate nor to underestimate amount of association commended rho as a solution to the skewed-data problem. However, that analysis did not rigorously test rho's ability to "track" differences in relationships. The present analysis adds to our knowledge of rho's performance with skewed data, and it may be necessary to retract the previous recommendation.

Another CA matrix comes from the previously discussed Viet Nam data (Section III.2 above). The escalation (ESC) and de-escalation (DE-ESC) scales can be arrayed by analysis model (bivariate Pearson r, eta, simultaneous multiple regression, and stepwise multiple regression), within the Eastern and Western U.S. populations.

Table III.4/2 shows a general tendency for the two bivariate models to agree better with each other than with the multivariate models. The multivariate models also tend, less strongly, to agree better with each other than with the bivariate models. In evaluating the models, Table III.4/2. The Correlation of Viet Nam Escalation and De-escalation Sentiment across Four Analysis Models: Pearson *r*, *Eta*, Stepwise Multiple Regression (SW), and Simultaneous Multiple Regression (MR) (CA Matrix, Profile Replication).

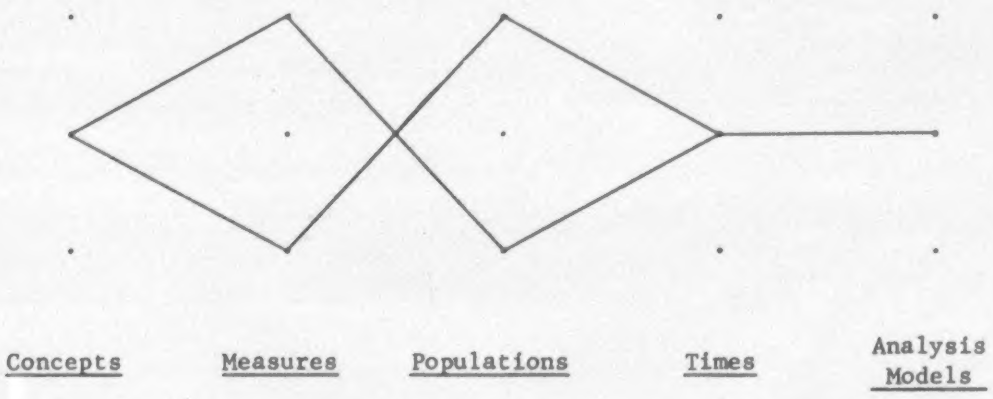
		Escalation				De-escalation			
		r	*SW*	*MR*	*eta*	*r*	*SW*	*MR*	*eta*
EAST:									
Escalation	*r*	-							
	SW	84	-						
	MR	79	89	-					
	eta	94	76	77	-				
De-escalation	*r*	-43	-25	-13	-34	-			
	SW	-19	-12	-07	-23	74	-		
	MR	-14	-10	-12	-16	69	96	-	
	eta	-35	-18	-07	-29	88	73	66	-
WEST:									
Escalation	*r*	-							
	SW	88	-						
	MR	92	90	-					
	eta	94	80	81	-				
De-escalation	*r*	20	15	03	30	-			
	SW	01	-02	-09	13	85	-		
	MR	00	-04	-11	13	77	94	-	
	eta	36	30	18	46	96	82	76	-

it is significant that only the multivariate models brought out the expected negative correlation between ESC and DE-ESC in the West, where the concept's profiles are not as distinctive as in the East.

One reason why the correlations between ESC and DE-ESC are not more negative, and why they are actually positive in the bivariate models in the West, is that certain profile elements show them rising and falling together. In particular, respondent's education (one of the predictors whose correlation with the concept makes up part of each profile) predicts both ESC and DE-ESC in the same direction, because it was the better-educated respondents who had any opinion on these issues. Other predictors in each profile predict both ESC and DE-ESC high, partly because of their correlations with education. In multivariate analysis, although partialling does not negate the correlation of education itself with ESC and DE-ESC, other predictors are free of the influence of education, and the expected negative correlation between ESC and DE-ESC emerges.

III.5.

The Measures by Populations (MP) Convergent-Discriminant Matrix



Validation of a measure across populations is a necessary condition for external validation of the concept measured. If the concept is operationalized via a single measure, these validations are one and the same procedure.

As in any matrix with population variation, profile matching is the only approach that provides pairwise observations for correlation. The profile points can be times, or population subgroups, or other measures.

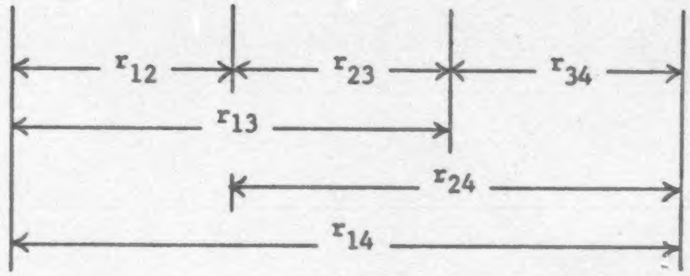
Most illustrations of the MP matrix resemble those of the CP matrix since concepts are often equated with single measures. An exception would be any "concept-free" measurement procedure, like the Semantic Differential. Data from such instruments, organized to fill an MP matrix (i.e., summed over concepts for reliability), would not fill a CP matrix.

It is appropriate to use the MP matrix to test alternative measures of a concept for stability across populations in the same analysis steps that require the CM matrix for construct validation. If there is greater doubt about stability than about construct validity, the MP analysis could be completed first.

As an example of an MP analysis with time as the replication dimension, suppose that two measures of the concept political unrest will converge in measuring changes over time in each of two populations and are not responsive only to international cycles such as economic depression. Assume that data have been collected on deaths per million from domestic noncriminal violence (M1) and protest party votes cast per million (M2) in Japan (P1) and France (P2).

Figure III.5/1. Convergent-Discriminant Matrix: Measures by Populations (MP)

		Measure 1 Population 1	Measure 1 Population 2	Measure 2 Population 1	Measure 2 Population 2
Replication over concepts or times, also other measures and population	1	x_{111}	x_{121}	x_{211}	x_{221}
	2	x_{112}	x_{122}	x_{212}	x_{222}
	3	x_{113}	x_{123}	x_{213}	x_{223}
	.	$x_{11.}$	$x_{12.}$	$x_{21.}$	$x_{22.}$
	n	x_{11n}	x_{12n}	x_{21n}	x_{22n}

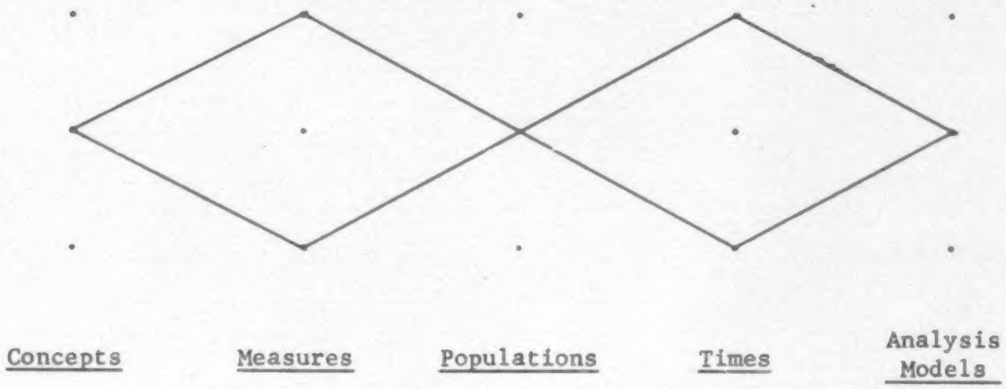


		M1		M2	
		P1	P2	P1	P2
M1	P1	r_{11}			
	P2	r_{12}	r_{22}		
M2	P1	r_{13}	r_{23}	r_{33}	
	P2	r_{14}	r_{24}	r_{34}	r_{44}

If r_{12} and r_{34} are large relative to r_{13} and r_{24}, then M1 and M2 are "tracking" changes extraneous to political unrest, such as an economic cycle. However, if the reverse is true and r_{13} and r_{24} are large relative to r_{12} and r_{34}, then convergence of the two measures within each country and discrimination between countries provides some basis for proceeding to construct validation via the CM matrix, although other artifacts remain to be checked.

III.6.

The Measures by Times (MT) Convergent-Discriminant Matrix

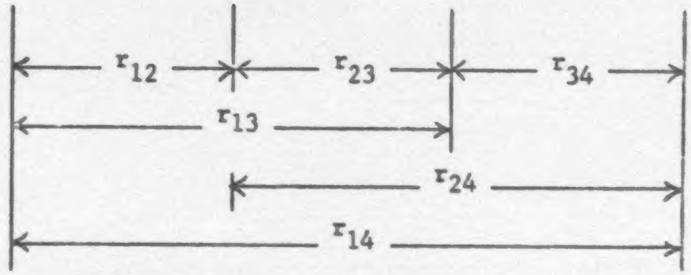


Perhaps the most important use of the MT matrix is that of assuring greater agreement between two measures of the same concept at the same time than between earlier and later administrations of one measure, assuming that the concept has different true levels at T1 and T2. Table III.6/1 shows that this condition is satified for A and B, two halves of a personality test, administered to college students as freshmen and then again as sophomores (Webster and Bereiter, 1967).

Although A correlates higher with A over time than it does with B over time (monomeasure-heterotime greater than heteromeasure-

Figure III.6/1. Convergent-Discriminant Matrix: Measures by Times (MT)

		Measure 1 Time 1	Measure 1 Time 2	Measure 2 Time 1	Measure 2 Time 2
Replication over concepts or populations, also other measures and times	1	x_{111}	x_{121}	x_{211}	x_{221}
	2	x_{112}	x_{122}	x_{212}	x_{222}
	3	x_{113}	x_{123}	x_{213}	x_{223}
	.	$x_{11.}$	$x_{12.}$	$x_{21.}$	$x_{22.}$
	n	x_{11n}	x_{12n}	x_{21n}	x_{22n}



		M1		M2	
		T1	T2	T1	T2
M1	T1	r_{11}			
	T2	r_{12}	r_{22}		
M2	T1	r_{13}	r_{23}	r_{33}	
	T2	r_{14}	r_{24}	r_{34}	r_{44}

heterotime), still the highest correlations of all are the heteromeasure-monotime -- A with B in T1 and A with B in T2.

The MT matrix can also be used if our primary interest is in time periods themselves, as ages or eras. Multiple measures of the "spirit" or the "style" of the 90's, the 20's, the 60's, etc., should show higher validity values (heteromeasure-monotime) than any other values in the matrix.

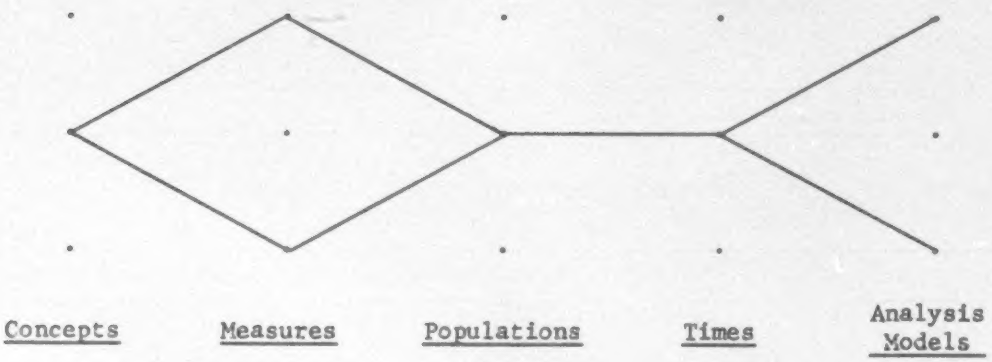
Table III.6/1. The Correlation of A and B Halves of a Personality Test across Time (MA Matrix, Replication over Persons).

		Freshmen (T1)		Sophomores (T2)	
		A	B	A	B
Freshmen (T1)	A	-			
	B	86	-		
Sophomores (T2)	A	73	67	-	
	B	67	69	83	-

Turning to a more esoteric MT application: if two measures of, say, the concept anxiety differed, intentionally or not, in their past versus future orientation (e.g., "Some recent events that I worried about" versus "Some events I worry about in the future"), then the future-oriented test in T1 ought to correlate more highly with the past-oriented test in T2 than vice versa. The example shows, at least, that even cross-lagged (heteromeasure-heterotime) correlations may be used in convergent mode for validation.

III.7.

The Measures by Analysis Models (MA) Convergent-Discriminant Matrix



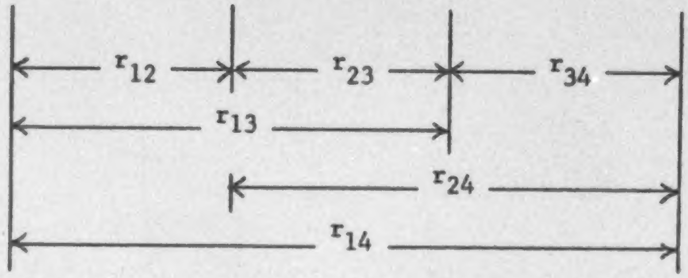
This matrix brings together the two methodological dimensions. It is the matrix for testing measure-model compatibility or robustness.

Table III.7/1 presents correlations based upon the "newspaper diffusion" and "radio diffusion" profiles once more. Correlations in this case have been averaged across the two concepts, so that only the measurement (estimated versus complete data sets) and analysis model (Pearson *r* versus Spearman *rho*) dimensions remain.

Table III.7/1 helps to answer the question, "Is it the measurement procedure or the analysis model that has the greated dampening effect on the correlation between one expression of a concept and another expression of the same concept?" Only newspaper-newspaper and radio-radio correlations were averaged together, since the

Figure III.7/1. Convergent-Discriminant Matrix: Measures by Analysis Models (MA)

		Measure 1 Analysis 1	Measure 1 Analysis 2	Measure 2 Analysis 1	Measure 2 Analysis 2
Replication over concepts, populations, or times, also other measrues	1	x_{111}	x_{121}	x_{211}	x_{221}
	2	x_{112}	x_{122}	x_{212}	x_{222}
	3	x_{113}	x_{123}	x_{213}	x_{223}
	.	$x_{11.}$	$x_{12.}$	$x_{21.}$	$x_{22.}$
	n	x_{11n}	x_{12n}	x_{21n}	x_{22n}



$$\begin{array}{cccccc} & & \multicolumn{2}{c}{\underline{M1}} & \multicolumn{2}{c}{\underline{M2}} \\ & & \underline{A1} & \underline{A2} & \underline{A1} & \underline{A2} \\ \underline{M1} & \underline{A1} & r_{11} & & & \\ & \underline{A2} & r_{12} & r_{22} & & \\ \underline{M2} & \underline{A1} & r_{13} & r_{23} & r_{33} & \\ & \underline{A2} & r_{14} & r_{24} & r_{34} & r_{44} \end{array}$$

cross-correlations (newspaper-radio and radio-newspaper) have an expected value of less than 1.00 -- namely, the true correlation between the profiles of the two concepts. The newspaper-newspaper and radio-radio correlations should equal 1.00; to the extent that they do not, either the measure or the analysis model is distorting one or both of the profiles. In other words, this matrix (although not the MA matrix in general) has unusual expected values of 1.00. Distance from 1.00 is an indication of the dampening effect of measure and/or model on the concept profiles.

Table III.7/1. The Correlation of Estimated and Complete Data Measurement Procedures across Pearson *r* and Spearman *rho* Analysis Models (MA Matrix, Profile Replication).

		Estimated		Complete	
		r	*rho*	*r*	*rho*
Estimated	*r*	-			
	rho	945*	-		
Complete	*r*	978	915	-	
	rho	962	976	958	-

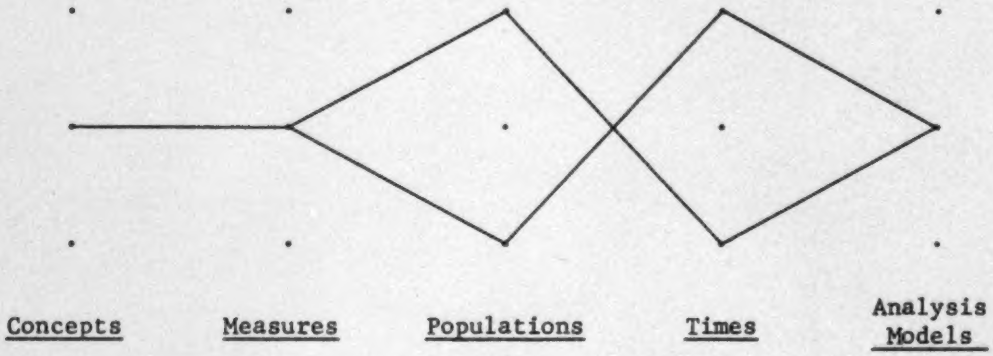
Clearly, the analysis model has the greater effect. Two measurement approaches employing the same analysis model correlate from .976 to .978. Different analysis models, whether or not they share the same measurement approach, correlate from .915 to .962.

---

*Correlation to three decimal places, decimal point omitted.

III.8.

The Populations by Times (PT) Convergent-Discriminant Matrix



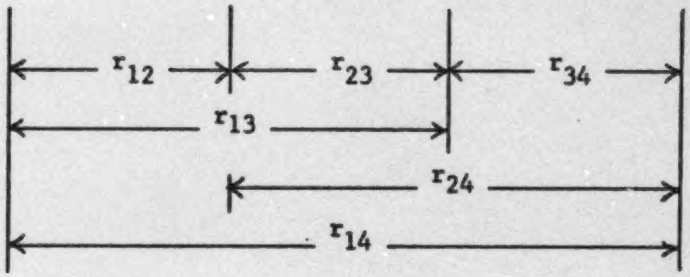
As stated above (Section III.3), the PT matrix supports the inference that one population (perhaps a single person) is changing over time toward the model or profile of attributes earlier exhibited by another population (or person). This CDM is so rich in potential application that we will list several hypothetical examples.

Example 1: Early and late in a political campaign, measure the "images" of two political candidates, using a measure-replicating instrument like the Semantic Differential (by "measure-replicating" we mean that correlations are computed between images across successive measures -- bipolar adjective scales -- that the images have in common). The images could be construed either as concepts or as person-perceptions; we choose the latter for this example, but of course the label is arbitrary and a CT matrix would also serve the purpose.

Just to flesh out the example, let P1 be Kennedy and P2 be Nixon. T1 and T2 are early and late in the course of the "Great Debates".

Figure III.8/1. Convergent-Discriminant Matrix: Populations by Times (PT)

		Population 1 Time 1	Population 1 Time 2	Population 2 Time 1	Population 2 Time 2
Replication over concepts or measures, also other populations and times	1	x_{111}	x_{121}	x_{211}	x_{221}
	2	x_{112}	x_{122}	x_{212}	x_{222}
	3	x_{113}	x_{123}	x_{213}	x_{223}
	.	$x_{11.}$	$x_{12.}$	$x_{21.}$	$x_{22.}$
	n	x_{11n}	x_{12n}	x_{21n}	x_{22n}



		P1		P2	
		T1	T2	T1	T2
P1	T1	r_{11}			
	T2	r_{12}	r_{22}		
P2	T1	r_{13}	r_{23}	r_{33}	
	T2	r_{14}	r_{24}	r_{34}	r_{44}

The correlations r_{12} and r_{34} show the stability of each candidate's image over time. The correlations r_{13} and r_{24} show similarity between candidates within each time period. The cross-lagged correlations r_{14} and r_{23} show the tendency of either candidate at T2 to approach the other candidate's image at T1 (conscious or unconscious modelling). There is anecdotal evidence from 1960 that Nixon tried to adapt his "television image" to narrow the "image gap" separating him from Kennedy.

Example 2: Measure similarity between the United States (P1) and Colombia (P2) on many dimensions (measures) of national development in 1940 and 1960 (T1 and T2). As before, the heteropopulation-monotime correlations r_{13} and r_{24} reflect similarity between the nations within each time period. The monopopulation-heterotime correlations r_{12} and r_{34} reflect stability in each nation's profile over time.

If the heteropopulation-heterotime correlation r_{14} is significantly large relative to r_{23}, then we infer that Colombia is becoming more similar to the U.S. over time than vice versa. If r_{14} is larger even than r_{34}, we might speculate on the extent of the development lag, since this finding tells us that Colombia in 1960 is more similar to the U.S. of 1940 than to the Colombia of 1940, implying that Colombia has come more than halfway along some continuum of change that separated it from the U.S. in 1940.

Example 3: This example breaks the mold of the 2x2 convergent-discriminant matrix. Measure "life style" in each of the 48 contiguous states in 1950 and 1960, using such Census indices as educational level, family size, deployment of the workforce, income level, etc.

Correlate all states in 1950 with all states in 1960, also all 1950 with 1950 and 1960 with 1960, taking the "life style" measures (transformed individually to z-scores) as the replication dimension. If, like the nations of the world, some states are highly developed while others are in an intermediate stage of development and others are relatively undeveloped, then we would expect a state in the intermediate development stage in 1960 to correlate more highly with the 1950 profile of a highly developed state than vice versa. Also, a state at the bottom of the development scale in 1960 would correlate more highly with a state at intermediate level in 1950 than with a highly developed state in 1950. For example, California may be becoming more like New York, and Iowa and Minnesota like Illinois, but, moving down the development scale, Colorado may resemble the Iowa of the earlier time period, the later Mississippi may resemble the earlier North Carolina, etc.

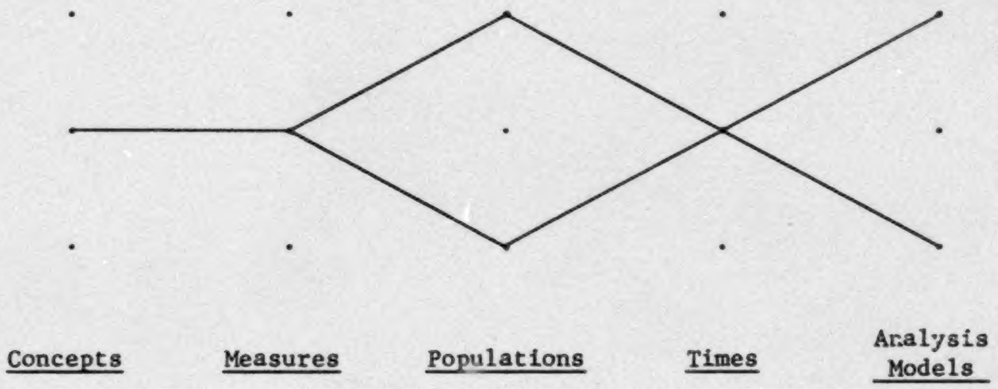
The promise of this kind of PT matrix, using profiles of development measures as the replication dimension, is that patterns of identification and modelling may be clarified. With the forthcoming Russet et al. time-series international development data, it should be possible to examine the similarity of nations within and between time periods in PT matrices. Taking P as Person, the same analysis would be rewarding in the study of individual growth and modelling. Of course, the adequacy of the set of measures, according to the pattern that we say is being modelled, will always be open to question.

Example 4: A multimeasure instrument like the Semantic Differential is used to measure attitudes toward communism among a population with grade school education (P1) and a population with college education (P2). The two time points are 1950 and 1970.

As in the previous examples, correlations show stability across time within each population, similarity between populations within each time period, and convergence of one population on the norm, or more stable pattern, established by the other.

III.9.

The Populations by Analysis Models (PA) Convergent-Discriminant Matrix

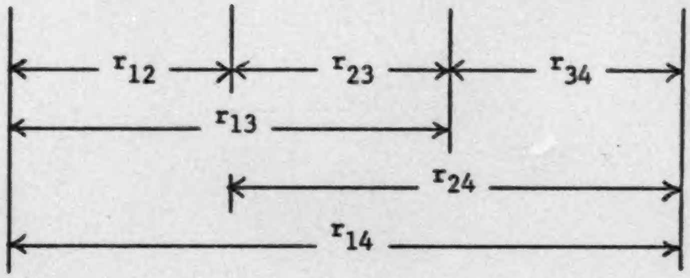


External validation, the successful generalization of a finding across populations, has long been sought in social research. Generally, however, the analysis model that generated the initial finding does not change. Outcomes of different models are not compared; hence whatever biases the model imposes on the data are likely to be found in both populations.

We are alluding to two levels of analysis model assessment. First, the finding should indeed appear in a new population. That is, the singular analysis model (and all antecedent operations) should prove out. Second, however, the analysis model should surpass other models on such criteria as over all agreement across populations, agreement with

Figure III.9/1. Convergent-Discriminant Matrix: Populations by Analysis Models (PA)

		Population 1 Analysis 1	Population 1 Analysis 2	Population 2 Analysis 1	Population 2 Analysis 2
Replication over	1	x_{111}	x_{121}	x_{211}	x_{221}
	2	x_{112}	x_{122}	x_{212}	x_{222}
concepts, measures, or times, also other populations	3	x_{113}	x_{123}	x_{213}	x_{223}
	.	$x_{11.}$	$x_{12.}$	$x_{21.}$	$x_{22.}$
	n	x_{11n}	x_{12n}	x_{21n}	x_{22n}



		P1		P2	
		A1	A2	A1	A2
P1	A1	r_{11}			
	A2	r_{12}	r_{22}		
P2	A1	r_{13}	r_{23}	r_{33}	
	A2	r_{14}	r_{24}	r_{34}	r_{44}

the consensus of other models, freedom from detectable bias, and the theoretical meaningfulness of outcomes.

Table III.9/1 contains still another view of the Viet Nam data that we presented above in Sections III.2 and III.4. Here four analysis models are correlated across two populations, with values obtained for the ESC and DE-ESC subanalyses averaged together.

Table III.9/1. The Correlation of Four Analysis Models: Pearson r, Eta, Stepwise Multiple Regression (SW), and Simultaneous Multiple Regression (MR), across Two Populations (PA Matrix, Profile Replication).

		East				West			
		r	SW	MR	eta	r	SW	MR	eta
East	r	-							
	SW	79	-						
	MR	72	92	-					
	eta	91	74	71	-				
West	r	72	61	66	75	-			
	SW	53	55	59	63	81	-		
	MR	56	59	64	62	84	92	-	
	eta	72	63	68	74	95	81	78	-

Like the MA matrix discussed in Section III.7 above, this PA matrix strains for 1.00 values in every cell. There is no basis for discrimination; failure of complete agreement reflects on the analysis models, on true differences between the populations, and on error variance adding "noise" to each analysis outcome.

There may be more subtle ways to compare the performance of the models across populations, but median values tell us that stepwise regression agrees with other models and itself most poorly (median = 63), followed by simultaneous multiple regression (68), Pearson _r_ (72), and _eta_ (74).

A PA matrix could also be filled only with discrimination values. Instead of correlating ESC with ESC and DE-ESC with DE-ESC (i.e., the convergence values) across models and populations, ESC from P1 could be correlated with DE-ESC from P2 and vice versa, with strong negative correlation as the goal. When only discrimination values (heteroconcept-heteropopulation) are tabulated, the two multivariate procedures show stronger discrimination than Pearson _r_ and _eta_. This is a special instance of the finding that, both across and within populations, it is simultaneous multiple regression that best discriminates in all heteroconcept cells.

Whether a researcher prefers to maximize the convergent _or_ discriminant power of an analysis model depends on his theory, the concepts in question, and his tendency to emphasize convergent or discriminant logic in validation. In this example, the bivariate models maximize convergence and therefore serve the traditional criterion of construct validation. The multivariate models maximize discrimination; although their monoconcept (convergence) values are not as high as those of the bivariate models, their heteroconcept (discrimination) values bring relationships between concepts into sharper relief.

Another example of PA application would be the comparison of factor analysis or cluster analysis solutions across populations. Factor analysis with "free-floating" rotation is notoriously difficult to generalize across populations. Often not even the same factors emerge; if they do, patterns of loadings shift. Simple nonparametric cluster analysis (Fruchter, 1954) has similar external validity shortcomings; the solution is inordinately affected by high correlations that may not appear in a matrix for another population.

The robustness of factor analysis or cluster analysis across populations can be tested in a PA matrix. For example, cluster analyses are performed on the same variables from two populations, say Viet Nam East and Viet Nam West. The "belongingness" criterion of cluster analysis causes the solution to terminate at _n_ clusters when cluster _n+1_ lacks internal cohesion relative to average association in the parent correlation matrix. This criterion causes different cluster analyses to terminate with different numbers of clusters. We can hypothesize that Viet Nam East yields a five-cluster solution while Viet Nam West yields a four-cluster solution. For purposes of the test, we would cut the five-cluster solution back to four clusters, for the sake of equivalence across populations.

Eliminating a cluster does not leave variables "orphaned", because the cluster analysis program has been modified to compute cluster loadings for every variable on every cluster, whether or not the variable belongs to the cluster. This modification produces an association matrix of _clusters by variables_ that is analogous to a factors by variables matrix of loadings. Even if a variable cannot win membership in one of the first four clusters, it still has a status -- a loading -- on each of the four clusters (the loading is simply its average association with all variables belonging to the cluster).

Once we know that the two cluster analyses are comparable with four clusters each, we can compute factor analyses to yield just four Varimax-rotated factors each. Then it is necessary to match clusters with clusters, clusters with factors, and factors with factors across the four solutions (in this example: cluster analysis of Viet Nam East, cluster analysis of Viet Nam West, factor analysis of Viet Nam East, and factor analysis of Viet Nam West). The matching step is necessary. Factor I in one solution might correspond to Factor III in the other solution and to Cluster II and Cluster IV in the two cluster solutions. Correspondence is based logically on overlap in content, but a simple operational definition involves direct correlation between the columns of loadings (i.e., if Factor I in one solution corresponds to Factor III in the other solution, the simple correlation of these columns of loadings will be higher than Factor I's correlation with any other factor in the other solution and vice versa). When matching is completed, total patterns of loadings can be correlated across models and populations to fill a PA matrix, and the heteromodel-monopopulation and monomodel-heteropopulation correlations can be used to evaluate model performance.

An empirical example of such use of the PA matrix is provided by Norman (1963). A five-trait personality test (see Section III.1) was administered to several samples of students. Five-factor solutions were computed independently for each sample, and convergence of solutions across populations was tested with the Kaiser factor similarity procedure (1960).

Coefficients of factor similarity ("strictly interpretable as correlation coefficients" -- Norman, 1963, p. 580) are reported for two samples in Table III.9/2. Although monopopulation coefficients are missing, enough data are reported to show impressive convergence on the major diagonal and discrimination in off-diagonal cells. Norman states that, of the 75 major-diagonal coefficients computed between all pairs of samples, none was lower than .888, and that, of the 300 off-diagonal coefficients, none was higher than .304. Thus the five-trait factor structure holds up well across replications.

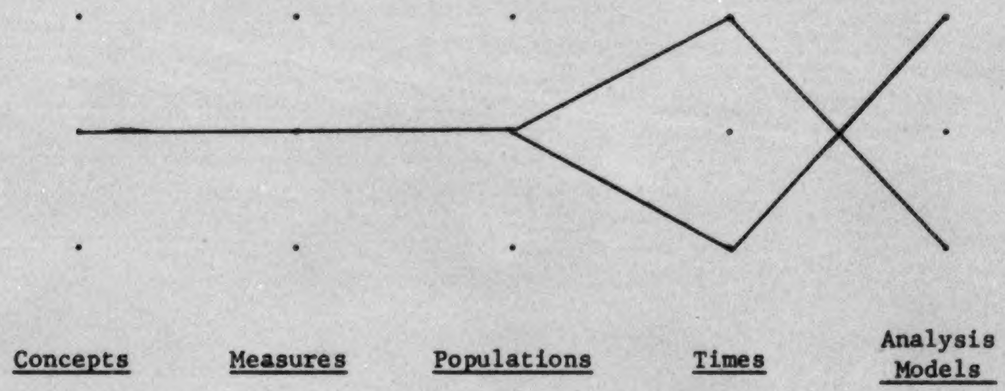
Table III.9/2. The Similarity of Factor Structures from the Same Personality Test Administered in Independent Samples (Partial PA Matrix*, Replication over Persons).

		Sample C				
		Factor I	Factor II	Factor III	Factor IV	Factor V
	Factor I	990	-023	-136	-009	-029
	Factor II	056	964	256	006	-049
Sample D	Factor III	110	-264	910	030	-300
	Factor IV	004	002	-037	999	-014
	Factor V	068	-034	295	024	952

*The two monopopulation triangles are not reported in the original article (Norman, 1963).

III.10.

The Times by Analysis Models (TA) Convergent-Discriminant Matrix

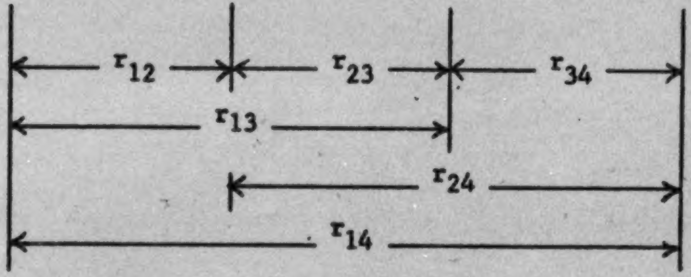


We have been hinting at one theme in many discussions of the three substantive choice-points -- concepts, populations, and times. It is a theme of first-and-last concern with relationships among concepts for the sake of simpler, stronger, and broader theories of behavior. If, as we began this paper, the product of science is reliable knowledge, then we can paraphrase ourselves: the product of science is an account of what has already happened and a prediction of what will happen, in particular to show that one phenomenon is a consequence of another (or of many others) across time and across aggregates of people. A study would be weakened, but not destroyed, by poor coverage of times and populations. Without concepts, however, there is no study.

Therefore the use of a matrix like TA or PA must be judged by the information it feeds back to our interpretation of concept relationships. We have no use for the ipsative finding that an analysis model performs differently from one time period to another unless that finding can be related to concepts of interest.

Figure III.10/1. Convergent-Discriminant Matrix: Times by Analysis Models (TA)

		Time 1 Analysis 1	Time 1 Analysis 2	Time 2 Analysis 1	Time 2 Analysis 2
Replication over concepts, measures, or populations, also other times	1	x_{111}	x_{121}	x_{211}	x_{221}
	2	x_{112}	x_{122}	x_{212}	x_{222}
	3	x_{113}	x_{123}	x_{213}	x_{223}
	.	$x_{1i.}$	$x_{12.}$	$x_{21.}$	$x_{22.}$
	n	x_{11n}	x_{12n}	x_{21n}	x_{22n}



		T1		T2	
		A1	A2	A1	A2
T1	A1	r_{11}			
	A2	r_{12}	r_{22}		
T2	A1	r_{13}	r_{23}	r_{33}	
	A2	r_{14}	r_{24}	r_{34}	r_{44}

Two kinds of analysis models can usefully be checked in a TA matrix. There are, first, the statistical association models that we have examined before -- r, eta, etc. (although regression models can be used to specify the nature of a relationship through a prediction equation as well as amount of covariance, we have been emphasizing the latter use). There are, second, prediction models more directly derived from theory, such as an exponential curve of diffusion, a logarithmic "decay" curve, a logarithmic rank-frequency relationship in language, etc. In both cases, we want to know whether the phenomenon, or our measure of it, has so changed over time that a previously appropriate model no longer fits.

In practice, we fill a TA matrix in the same way as a PA matrix. Since we are dealing with analysis outcomes, some kind of profile matching is necessary except in the rare case when both data sets come from the same populations or persons and analysis outcomes can be reflected back to P-units in the form of factor scores, residuals, etc.

IV. THE ROLE OF THE CONVERGENT-DISCRIMINANT MATRIX IN CAUSAL ANALYSIS: FORETHOUGHTS

Until now the recurring theme of this monograph has been that the researcher can validate his methodology at each decision-point. The strength or weakness of internal relationships (made apparent in the appropriate CDM matrix) will indicate the validity of his research.

However, if one chooses to work with the substantive matrices, one finds that there are certain external issues that must also be taken into consideration. The relevant matrices are the CP matrix, the CT matrix, and the PT matrix. Two of the three are novel; however, the CT matrix includes the tradition of the cross-lagged panel correlation (CLPC) used in causal inference.

An incomplete network of studies and techniques has developed around the notion of imputing causal relationships to variables in non-experimental data. As is true of networks, some of the lines of work represent parallel efforts extended in time without knowledge of each other. Some of the lines do cross. But only occasionally is there a line that intersects several techniques. Many of the studies continue to show a single-minded approach to causal analysis. After a review of the literature in this network, we find several issues which will be discussed under two major headings -- the language of causality and the techniques of assessing causality. Some of the issues raised in this chapter cannot be adequately answered. So, for now, let the asking of questions serve as a next step in the progress of causal inference.

IV.1. The Language of Causality

The philosophy of science influences our notions about causality. Philosophers such as Ernest Nagel and Herbert Feigl have centered our thinking on four main views of causation which continue to be used, either implicitly or explicitly, by many scientists. These are: (1) necessary and sufficient causation, (2) temporally sequential causation, (3) spatially contiguous causation, and (4) asymmetrical causation. Let's examine each of the four in an effort to determine its applicability to non-experimental causal inference.

Nagel (1961) lists the necessary and sufficient condition as the first aspect of a causal relationship. If the cause is present, then the effect will also occur. In the sense of a causal law, A, the cause, must be the necessary and sufficient condition for the presence of B, the effect. If A is only the necessary condition for B, then B will always be found in the presence of A, but A can appear in circumstances of non-B. If A is only the sufficient condition for B, then B will always be found in the presence of A, but B can also appear in circumstances of non-A. The enumeration of all possible combinations of conditions looks like this (with illustrative data supplied):

Table IV.1/1

A is necessary and sufficient for B

	A+	A-
B+	100	0
B-	0	100

Table IV.1/2

A is necessary for B

	A+	A-
B+	60	0
B-	40	100

Table IV.1/3

A is sufficient for B

	A+	A-
B+	100	40
B-	0	60

Table IV.1/4

A is partly necessary and/or partly sufficient for B

	A+	A-
B+	60	40
B-	40	60

The first three possibilities require one or more zero cells. Given measurement error, it is unlikely that we will ever find empirical examples of these ideal zero cells. And in the event that we do find these cells, we must reckon with the possibilities that in the next sample we draw, they will not be zero. This argument about the use of necessary and sufficient conditions parallels Karl Popper's notion of falsifiability (1961). For just as we can never prove our theories, we can never prove a necessary and/or sufficient condition. All we can say is that one particular sample or test did not falsify our assumption. Blalock (1961) emphasizes some of the same problems with necessary and sufficient conditions and then adds another. Let us elaborate his argument, which deals with dichotomies. Two preliminary statements need to be made. First, necessary and sufficient terminology is only applicable to dichotomous attributes. We could argue that if two continuous variables, X and Y, correlate 1.0, one is necessary and sufficient for the other. But we would not know which one is the cause and which is the effect. That would seem to lead us back to the old saying "correlation does not prove causation." However, the case is such that we rarely (if ever) have a correlation of 1.0, and any correlation less than unity makes talk of necessary and sufficient conditions seem inappropriate.

If we wanted to have a causal language for non-experimental data that included this terminology, we could argue that most continuous variables can be dichotomized. The question now becomes, "Where do we make the cut and what happens when we do?" And this leads us to our second point that where one dichotomizes a variable determines whether the relationship appears to be both necessary and sufficient, only necessary, or only sufficient. Figure IV.1/1 illustrates the plot of two highly correlated variables that have been evenly split. The adjoining table (IV.1/5) shows how this dichotomization leads one to infer that X is the necessary and sufficient cause of Y. Figure IV.1/2 shows a split that changes X so that 75 per cent is on the left side and 25 per cent is on the right side, and that changes Y so that 75 per cent is below the line and 25 per cent is above it. When these values are tabled (Table IV.1/6), we conclude that X is a necessary but not sufficient cause of Y. The third figure (IV.1/3) shows a dichotomization for X that puts 50 per cent on each side and for Y that puts 75 per cent above the line. Table IV.1/7 indicates that X is a sufficient but not necessary cause of Y. From these examples we see that, given the same underlying distribution with different choices of the dichotomization point, we might draw any of three conclusions about necessary and sufficient conditions of X for Y. There are many reasons why we might wish to dichotomize at different points. We might choose a median instead of a mean. Or, we might have data with five choice responses that when grouped differently (e.g. 1, 2, 3 versus 4, 5 or 1, 2 versus 3, 4, 5) give quite different results. Let it suffice for the present to say that dichotomization of continuous variables can give misleading causal indications and that the marginals of dichotomous variables must be carefully considered before discussing necessary or sufficient conditions.

Table IV.1/5

X is necessary and sufficient for Y

	X+	X-	
Y+	50	0	50
Y-	0	50	50
	50	50	100

Figure IV.1/1

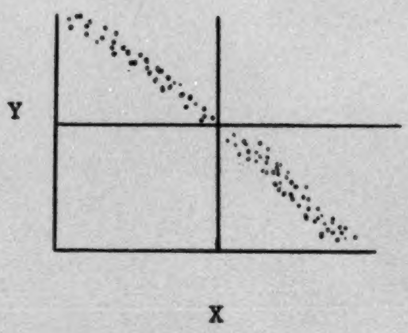


Table IV.1/6

X appears to be necessary for Y

	X+	X-	
Y+	25	0	25
Y-	50	25	75
	75	25	100

Figure IV.1/2

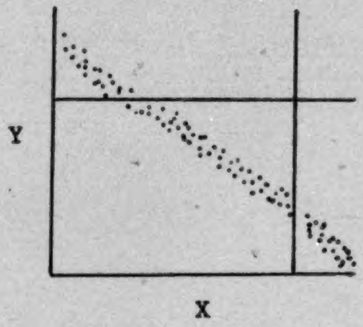
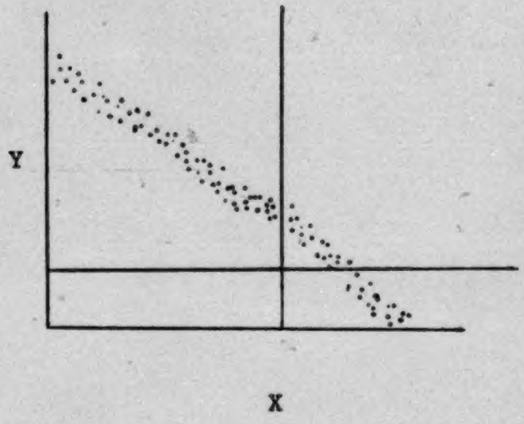


Table IV.1/7

X appears to be sufficient for Y

	X+	X-	
Y+	50	25	75
Y-	0	25	25
	50	50	100

Figure IV.1/3



Blalock suggests that we develop a language that can deal with continuous variables. Then, if no data assumptions are violated, and if we have an analysis model that requires us to, we can dichotomize the variable. Of course, we cannot make a continuous variable from a dichotomous variable.

Although Blalock makes no specific suggestions, we propose two terms for a causal language -- contributory and contingent. Contributory causation says that A partially determines B. For instance, Therapist warmth may contribute to (increase) Patient warmth. (Here the assumption might be that warmth is an indicator of Patient improvement.) In reality, C may work with A to cause B. However, since we have only measured A, and would like to specify to what extent A causes B, we find contributory to be a useful causal term. If we were to think of a dichotomy that indicated contributory cause, we would have a 2 x 2 matrix like the one previously labeled "partly necessary and/or partly sufficient." In other words, since it seems a contradiction to the necessary and sufficient terminology to say one variable is partly necessary for another, we say one variable contributes to another. The size of the contribution of A to B would be indicated by the size of the correlation coefficient.

Another type of causal relationship might be called the contingency specification. We might say that Therapist warmth causes Patient warmth if the Patient perceives that the therapy is helpful. Or, stated more generally, A causes B if C. A more complicated model might predict contingent causal relationships of the form A causes B if C, D, and E. Again a dichotomy makes it easy to demonstrate this type of relationship. Table IV.1/8 shows first the original relationship between Patient warmth and Therapist warmth, and second, the relationship when the helpfulness of therapy is held constant.

Table IV.1/8

	T+	T-
P+	40	10
P-	10	40

THERAPY HELPFUL

	T+	T-
P+	25	2
P-	0	23

THERAPY NOT HELPFUL

	T+	T-
P+	15	8
P-	10	17

This approach looks familiar to those acquainted with the Lazarsfeld-Kendall-Hyman elaboration technique called "specification" (Hyman, 1955). Its weakness of producing possible regression artifacts (Campbell and Clayton, 1961) should be understood by a potential user. The analysis model appropriate for interval data would be either a difference in correlation or partial correlation (for problems with the latter see Brewer, Campbell, and Crano, 1970).

The second aspect of causality that Nagel specifies is temporal sequentiality. Again we find ourselves taking the position that the present state of the art and the kind of variables we deal with make this a meaningless term. We have many situations in which the effect follows the cause. A jet breaks the sound barrier and the sonic boom creates sound waves that cause a window to break. Obviously the breakage comes after the boom, the effect after the cause. Or consider a persuasive message concerning the side effects of LSD that causes a change in attitude towards drug use. Again the cause (the message) precedes the effect (the new attitude). However, some variables may operate cyclically. We argue from a time-series model that the variable being called the cause is merely dependent on the particular time point chosen. For example, the Therapist and Patient warmth variables may cycle from one week to the next. In week 1 (W1) Therapist is warm and the Patient is not. In thinking about the session the patient realizes the discrepancy in moods, realizes the Therapist was responding in a warm way, and decides to show more warmth next session. In W2 the Therapist is warm and so is the Patient. In W3 the Therapist is critical, but the Patient continues to be warm. The Therapist then responds so that in W4 both score high on warmth. Clearly in this example we would say Patient warmth was the effect because it changed after the variable Therapist warmth. However, if one should measure the situation at W3 and W4, instead of W1 and W2, then Patient becomes the cause and Therapist the effect. A final example is drawn from an agricultural setting. Suppose, for instance, that rainfall measured at T1 appears to cause crop yields at T2. In reality, it was low rainfall and low yields that caused the seeding of clouds that then caused the rainfall. And in looking across the entire chain one sees that an event which was completed in the past (low yields) caused the future event (high yields).

To summarize the temporal sequentiality aspect, we note that the effect will always appear to follow the cause. However, if we have measured the two variables at an inappropriate time, or have an inappropriate causal variable, we may inaccurately describe the temporal sequence.

The third aspect, spatial contiguity, is somewhat harder to define for the behavioral sciences. As long as we are dealing with physical causation, space is a meaningful concept. However, although there must be some kind of contiguity, it is inadequate to use space as the carrier of causal relationships in the behavioral sciences. For instance, we could say that teachers liking for pupils and pupils liking for teachers occurs in a spatially contiguous situation -- the classroom. Similarly, the warmth of Patient to Therapist and of Therapist to Patient occurs across two spatially close chairs. However, this aspect of causality is almost trivial when the variables are attitudes or traits. The link between A and B could involve either storage units such as memory or communication units such as speech, print, and gestures. This idea follows closely Campbell's notion of "real storage processes" that account for the presence of "delayed-action causal sequences" (Rozelle and Campbell, 1969, p. 77).

Finally, Nagel says that the relationship must be asymmetrical. His example, coming from the physical sciences, says "that the passage of the spark through the mixture of gasses is the cause of their transformation into water, but the formation of the water is not the cause of the passage of the spark". Unfortunately, behavioral scientists do not have such a neat package. As Rozelle and Campbell found (1969), grades in a course are as likely to affect attendance as is attendance to affect grades. Or, who can say which is more reasonable, that Therapist feelings cause Patient feelings, or that Patient feelings cause Therapist feelings? In the behavioral sciences we are finding that not relying on *a priori* asymmetry allows us to formulate competing hypotheses. And it seems to the writers that, at present, this is the most we can do. We cannot really say that A causes B in a total sense. However, we can say, assuming we have controlled for spurious relationships (see IV.2.F), that of two competing hypotheses, A causes B and B causes A, one has the stronger empirical support.

IV.2. The Techniques

There is one dimension of causal analysis that has crept into some of the previous examples but has not been made explicit. This is the time dimension. As soon as one is not bound to a single time measurement but can have repeated measurements, one sees a variety of potentially powerful techniques for estimating causal influence. It is the time dimension that makes the CT and PT matrices possible.

One way of looking at the techniques most frequently used is to examine each in the historical order that it occurred. These techniques are the Lazarsfeld (1948) sixteenfold technique (updated in 1969), the Campbell and Stanley (1963) cross-lagged panel technique (CLPC), which was expanded by Rozelle and Campbell (1968) to include more competing hypotheses, the Yee and Gage (1968) techniques called frequency-of-shift-across-median and frequency-of-change-in-product-moment, and the Heise (1969) path analysis model for two and three wave data. Missing from this list is the Pelz and Andrews (1964) technique of cross-lagged differentials. Their work follows the tradition of the CLPC and is not reviewed separately.

Each technique has its own set of strengths and weaknesses. First, let's simply examine each approach, then evaluate the assumptions of each model. (Since this review can only highlight the approaches, the reader is urged to consult the original documents for more details.)

IV.2.A. Sixteenfold Table

The Lazarsfeld approach is intended to tell us, for instance, whether the Therapist's perception of the Patient at Time 1 (T1) is the cause of the Patient's perception of self at Time 2 (T2). As a hypothetical example, let's assume we have 100 therapists and 100 patients. After each of two sessions the therapists and patients fill out a questionnaire that ascertains the patient's perception of how warm the therapist was to him and the therapist's perception of how warm the patient was to him. The following table (IV.2/1) presents plausible results. Note that T+P+ means that both the therapist and the patient perceived themselves as warm during that session. This notation can be simplified to ++, +- (Therapist warm, Patient not warm), -+ (Therapist not warm, Patient warm), and -- (neither warm).

Table IV.2/1

Patient's Perception of Therapist and Therapist's Perception of Patient

		2nd Session				
		T+P+	T+P-	T-P+	T-P-	
	T+P+	26	5**	3**	8	42
1st Session	T+P-	10*	3	1	4*	18
	T-P+	4*	1	1	10*	16
	T-P-	5	2**	2**	15	24
		45	11	7	37	

*no. of adjustments **no. of losses of adjustmen

Each Therapist and Patient matched pair is considered adjusted or congruent[1] if it is either ++ or --. The eight cells in the table without asterisks represent cases where Therapist and Patient are congruent at both sessions or are incongruent at both sessions. The four single-asterisked cells are the cases where Therapist and Patient move into congruency. The cells with double asterisks are the cases where Therapist and Patient change to incongruency.

The asterisked cells are combined to evaluate the relative strength of the two plausible causal influences. The actual index, which is represented by:

$$I_{T,P} = \frac{8\left(\frac{\Delta_H}{N_H} + \frac{\Delta_V}{N_V}\right)}{N}$$

also eliminates the effects of sample size. For our example,

$$\frac{\Delta_H}{N_H} = \frac{\begin{vmatrix} 10 & 4 \\ 4 & 10 \end{vmatrix}}{28} \quad \text{and} \quad \frac{\Delta_V}{N_V} = \frac{\begin{vmatrix} 5 & 3 \\ 2 & 2 \end{vmatrix}}{12}$$

or $I_{T,P} = .21$.

Since this approach to causal inference does not give us a result in a meaningful metric, William Robinson derived a test of significance (reported in Lazarsfeld, 1948).

Lazarsfeld has recently revised his approach (1969a, 1969b). Although he still supports the use of a sixteenfold table, he now believes that his previous index of "mutual effect" was not sensitive enough to establish causal influence in the many situations where direction is not easily determined.

---

[1] Lazarsfeld has used the terms adjustment and congruency interchangeably.

In these two latest manuscripts, Lazarsfeld discusses several new indexes that have been developed by others. Among these approaches he seems to lean toward the work done by Boudon (1968). By reworking some of Boudon's equations, Lazarsfeld derives a bilinear model for estimating the parameters of his schema. It contains eight coefficients. Of these, four are quite important. Two give us what Lazarsfeld calls "triple covariance"[1]. He defines these two coefficients as "the rate by which the turnover of one variable changes as we move along the scale of the other variable ascertained the first time." The second pair of coefficients represent triple covariances "between one observation the first time and two observations the second time." He goes on to say that he ends up "with something similar to a conventional partial correlation coefficient. But it [the triple covariance] should not be confused with it and cannot be computed from second level co-variances.[2] We deal here with third level co-variances from which the second level ones have been partialed out."

In another manuscript, Lazarsfeld speaks highly of Coleman's (1964) qualitative model for estimating mutual effects. Lazarsfeld summarizes this approach:

> He assumed that if a person was, say, in the State +- there would be two forces to draw him respectively into State ++

---

[1] He uses the terms correlation and covariance interchangeably.

[2] He renames zero-order correlations and calls them second level correlations.

and --, two more forces acted on people in initial State -+. To this he added four random oscillations pulling people away from wherever they are. Coleman showed that these eight parameters could be estimated from an empirical sixteenfold table and he too compared them successfully with fitted reconstruction from the model.

For the purposes of this chapter no further details of the Coleman technique for analysis of a sixteenfold table are necessary. However, the interested reader will find an informative description in a manuscript by Hornik (1969).

One final quote from Lazarsfeld seems accurately to reflect his present emphasis.

> ...the mark of the root problem: how to unmix a sample of people characterized by a set of response patterns into "latent subgroups" within which all contingent cross-products vanish. (1969a)

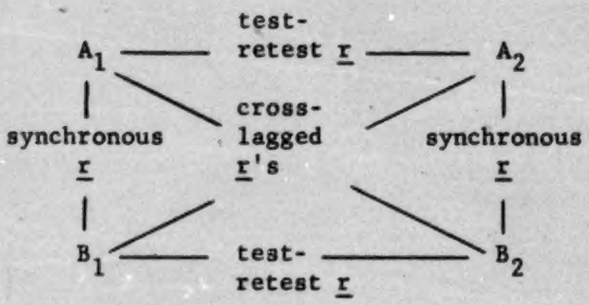
There were four problems with the original Lazarsfeld technique. First, even though there is a significance test, the results are not in a convenient metric. Second, it requires dichotomous data, or data that one is willing to dichotomize. Third, it is subject to regression when the marginals of either variable are skewed. And further, it cannot handle spurious relationships. Most of these objections still apply to his revised technique.

IV.2.B. Cross-lagged Panel Correlation

In an effort to compensate for these flaws, Campbell and Stanley propose the cross lagged panel technique (CLPC), using correlations.

There are several implications to this. There is a meaningful metric (the result will always vary between -1.00 and 1.00 with standardized tests to infer significance), continuous variables can be used, and regression artifacts can not influence the data analysis (although skewed distributions still can).

The pattern of correlations being examined will look like this:



The original CLPC procedure considered only the *source* of influence. It permitted the researcher to choose between the competing hypotheses *A* causes *B* and *B* causes *A*. If *A* is the cause, then the correlation of $A_{T1}B_{T2}$ is greater than the correlation of $A_{T2}B_{T1}$. A short time later, Rozelle and Campbell (1969) extended the list of rival hypotheses to four so that *source* and *direction* are examined. These are: (1) increases in *A* cause increases in *B* and decreases in *A* cause decreases in *B*, (2) increases in *A* cause decreases in *B* and decreases in *A* cause increases in *B*, (3) increases in *B* cause increases in *A* and decreases in *B* cause decreases in *A*, and finally (4) increases in *B* cause decreases in *A* and decreases in *B* cause increases in *A*. Rozelle and Campbell say these four are confounded in two pairs so that when $r\ A_1B_2 > r\ B_1A_2$, the joint effect of *A* increases *B* and *B* decreases *A* is greater than the joint effect of *B* increases *A* and *A* decreases *B*. To choose between the confounded pair, Rozelle and Campbell suggest comparing the cross-lagged correlations to a no-cause baseline, which ideally is reflected in identical zero values for the synchronous correlations. When we have data that does not show these zero values, then some kind of a baseline must be estimated. At present, Campbell suggests using the following procedure:

1. correct the test-retest correlations for attenuation (assumes knowledge of reliability for variables A and B);
2. obtain the geometric mean of the two corrected test-retest correlations;
3. multiply the geometric mean by the synchronous correlation (if the two synchronous values are not identical then take the simple average of the two before multiplying with the geometric mean).

Kenny (1969) has documented and refined the no-cause comparison baseline. He makes explicit the assumption behind such a model. However, without going into the specifics we can say that the effect of this procedure is to obtain a no-cause baseline value that takes into account reliability attenuation and temporal attenuation factors. This no-cause baseline gives us the value appropriate to the situation in which a third variable C is the cause. Then if our two values differ significantly from the no-cause line we can infer which causal influence is greater. For example, if A_1B_2 is greater than B_1A_2 and also greater than the no-cause baseline, we would conclude that A causes B.

Although some of the problems with the CLPC are still being worked on (e.g. the most appropriate baseline), there are certain basic assumptions about the model that must be noted before data can be applied. The two main assumptions to keep in mind are: (1) the synchronous correlations should be identical except for sampling errors, and (2) the test-retest correlations should show attenuation over time. The most serious of these two is the first, for it is possible to conceive of data sets from which one would expect synchronous correlations that increase or decrease over time or show regular cycles of increasing and decreasing.

This particular aspect of causal analysis is made explicit by Howard and Krause (1969). They reason that the causal situation can be one in which a state is maintained or in which a change is initiated. These conditions become four causal propositions: (1) Static-Static, where a state or level of (for example) Therapist's warmth causes or maintains a level of Patient warmth. This might be appropriate during the middle of therapy when the Therapist and Patient have gotten to know each other fairly well and when the Therapist can maintain the Patient's warmth level by maintaining his own level. (2) Static-Dynamic, where a state of the Therapist is the cause for a change in the Patient. Consider the situation early in psychotherapy where the patient is expecting many changes. Here the Therapist may maintain a consistent level of warmth while the Patient's warmth may increase (or decrease if the sessions are not going as well as the Patient would like). (3) Dynamic-Static, where a change in the state of the Therapist is the cause for a constant state of the Patient. An example of this would be when a Patient has been resisting treatment and the Therapist finds that he has to increase the amount of his warmth just to maintain the Patient at a constant level. (4) Dynamic-Dynamic, where a change (increase or decrease) in the Therapist is the cause of a change (increase or decrease) in the Patient.

If we have measurements at only two time points, then we must deal with a Static-Static or a Static-Dynamic model. However, if we have three or more measurements, we open up the possibility of the other two models, Dynamic-Static and Dynamic-Dynamic. In other words, if we know the Therapist's score at only T1 and T2, we can only talk about the static T1 score and its influence on P2 (we do not know if the T1 score has increased, decreased or remained the same from T0 to T1). Unless we can collect measurements on many occasions, we may have to be content with two time points and consequently only two possible models. Even if the relationship is Dynamic-Static or Dynamic-Dynamic, we will not have data to ascertain if the causal variable is increasing or decreasing.

Returning to the CLPC and reconsidering it now within the framework of the CDM, we see how more than two variables can be compared at one time (Table IV.2/2). The correlations in the monotime-heteroconcept triangles are the synchronous values. The heterotime-monoconcept diagonal is filled with the auto correlations. And the heterotime-heteroconcept cells contain the cross-lagged correlations. This means that in the example below we can check the synchronous correlations for Literature (Lit) with Word Functions (WF), Reading Comprehension (RC) and Abstract Reasoning (AR) at both time points, take into account each variable's auto correlation, and then look down the left column and across the T2 Literature row (as indicated by the arrows) to evaluate each of three different variables cross-lagged with Literature.

Table IV.2/2

		T1				T2			
		Lit	WF	RC	AR	Lit	WF	RC	AR
T1	Lit	Rel.							
	WF	Syn	Rel.						
	RC	Syn	Syn	Rel.					
	AR	Syn	Syn	Syn	Rel.				
T2	Lit	Auto	CL	CL	CL	Rel.			
	WF	CL	Auto	CL	CL	Syn	Rel.		
	RC	CL	CL	Auto	CL	Syn	Syn	Rel.	
	AR	CL	CL	CL	Auto	Syn	Syn	Syn	Rel.

In moving causal inference into the CDM framework we find we have not eliminated our problems, but instead have only altered the methodological maze somewhat. As Rubin (1970) has pointed out, analysis of the M^3 (and, in fact, all CDM matrices) can follow two paths -- numerical analysis procedures and model parameter estimation procedures. He goes on to say that the main emphasis of the original M^3 and the procedure for estimating the influence of method and trait factors proposed by Campbell and O'Connell (1967) follow the path of numerical analysis procedures.

Before continuing to other causal techniques, let us discuss in greater detail some of the analytical procedures of the one technique, CLPC, that best represents the spirit of this monograph. This side path will include a brief review of two articles and a dissertation proposal that are relevant to the evaluation of CDM matrices. The first two, Campbell and O'Connell (1967) and Jackson (1969), follow the numerical analysis procedures path. The proposal outlines a potentially usable model parameter estimation procedure.

Campbell and O'Connell explore the effects of a method factor on the M^3. A priori they specified the three best models: (1) an additive factor component, (2) a non-additive (inversely proportional) component, and (3) a multiplicative factor component. This means, using the additive factor as our example, that monomethod-heterotrait correlations would be higher than the heteromethod-heterotrait values because the former share a method component that has artifactually inflated the correlations by a constant value.

The authors create scatter plots for a variety of data sets, and use the slope as a measure of the "shape of the scatter independent of the degree of correlation" (p. 416). An examination of the results indicates to the authors that the multiplicative model is probably the most accurate.

Of particular interest here is that Campbell and O'Connell include data that would, in our terminology, fill a CT matrix. They discuss the presence of time factor and conclude that it might correspond to a method factor. They state:

> The degree of attenuation of the heterotemporal values is perhaps described by the difference between synchronous or internal consistency reliability measures and the test-retest reliability values which occupy what is here treated as the "validity" diagonal. The monotemporal matrix might just be the hetero-temporal matrix corrected for such temporal attenuation. Notice that the traditional correction for attenuation acts in a multiplicative way, increasing low correlations less than high ones. (p. 423-424)

Although the article is tentative in much that it says, it indicates the need for a way to interpret the M^3, or more generally, the CDM.

Another step along the numerical analysis procedures path has been taken by Jackson (1969). He proposes a multimethod factor analysis that allows one to evaluate convergent and discriminant validity. He first makes several points that are relevant to the evaluation of all CDM's, not just our CT matrix. He finds fault with "eye-ball" techniques because they emphasize comparisons among individual correlations rather than examination of the structure of the entire matrix. He also points out that the Campbell and Fiske criterion of heteromethod-monotrait values exceeding the monomethod-heterotrait correlations is particularly untenable. He is aware, of course, that Campbell and O'Connell have worried about this issue, and he agrees with their conclusion that "correlated errors of measurement are more likely to occur on a single occasion of testing and to operate strongly in the monomethod section...". (p. 34).

To meet the agreed-upon need to separate trait and method factors, Jackson discusses two techniques. There is Tucker's multimode factor analysis (1966), which "seeks to identify, in the trait mode, a recurring common factor structure across more than one method of measurement rather than to ascertain a single test's or trait's convergent and discriminant validity across methods of measurement." (p. 35).

Then there is the Jackson multimethod factor analysis. Jackson says that solutions arrived at through traditional factor analysis often have spuriously simplified structure because method variance confounds two or more traits. Trait and method factors must be separated (orthogonalized).

But before factor analyzing the matrix we need to bear in mind that each consists of several submatrices and that the monomethod matrices contain common variance that derives from method and from heteromethod trait variance, while the heteromethod matrices include only the "trait variance which is present across methods of measurement." (p. 39). To eliminate the intruding method variance Jackson orthogonalizes the monomethod matrices by substituting identity matrices. He says (p. 40): "In addition to permitting separation of method from trait variance, this prior orthogonalization procedure yields multitrait-multimethod matrices" that will result in a large number of factors that are fairly independent of method variance.

A possible limitation with the procedure is that it appears to work best with highly dissimilar methods and that it does not have a separate way of identifying method factors. In other words, the factor solution is not biased by the influence of an overall method factor, but if we wanted to determine the size or extent of effects of such a method factor, we could not get the needed information from the Jackson technique.

Rubin's dissertation proposal, relying on Bock and Bargmann (1966), sets up a framework for analysis of the M^3. He proposes to analyze covariance structures via statistical model building procedures. There are three by-products of this approach: (1) a X^2 goodness of fit test, (2) model parameter estimates, and (3) a statistical assessment of the estimates.

The M^3 "implies a classification of test items, usually a simple factorial" (Bock and Bargmann, 1966). Consequently, Rubin says that the _a priori_ relationship between factors (whether these be measure and concept factors, or time and concept factors, or a variety of other factor types) must be orthogonal. This is similar to Jackson's position but contradicts Boruch, who has argued that factors be oblique (Boruch and Wolins, 1968, Boruch, *et al.*, 1970). He suggests handling error variances and scale factors through the use of maximum likelihood estimation techniques. The derived general model for test scores is analogous to that for multivariate analysis of variance.

IV.2.C. Frequency-of-Change-in-Product-Moment

Let us continue with the examination of some causal techniques currently in use. Yee and Gage (1968), working at about the same time as Rozelle and Campbell, came to a similar conclusion regarding the importance of assessing both source and direction. However, their approach is closer to the Lazarsfeld technique in that they are concerned with shifts toward or away from congruence. Of the Rozelle and Campbell hypotheses: (1) increases in *A* cause increases in *B* and decreases in *A* cause decreases in *B*, *and* (3) increases in *B* cause increases in *A* and decreases in *B* cause decreases in *A* would be considered *congruent* changes; while (2) increases in *A* cause decreases in *B* and decreases in *A* cause increases in *B*, *and* (4) increases in *B* cause decreases in *A* and decreases in *B* cause increases in *A* would be considered *incongruent* changes.

Two techniques for evaluating these shifts from Time 1 to Time 2 are discussed by Yee and Gage. The first, called the *frequency-of-shift-across-median*, trichotomizes each variable so that one is considering scores above the median, at the median, and below the median. Then the researcher categorizes his data into a 9x9 matrix that shows the two possible sources of influence (e.g., Therapist and Patient) in all combinations of the trichotomized variable (e.g., Therapist, above the median score for warmth at T1 paired with above the median score for warmth at T2, and Patient, above the median score for warmth at T1 paired with below the median score on warmth at T2, etc.). From this it is possible to determine the causal influence -- Therapist, Patient, or uncertain (e.g., no change, change in same direction) and whether the T1 to T2 change is congruent, incongruent, or uncertain. When summarized, these categories look like this:

Source of Influence	Direction of Influence		
	Congruity	Incongruity	Uncertain
Therapist	TC	TI	TU
Patient	PC	PI	PU
Uncertain	UC	UI	UU

These nine summary variables are then regrouped in three ways and each is tested with chi-square. These regroupings are: (1) data pairs showing Therapist influence toward either congruity or incongruity occur more often than data pairs showing Patient influence toward either congruity or incongruity (this would be shown in the case where $TC + TI + TU > PC + PI + PU$), (2) data pairs showing Therapist influence toward congruity is greater than Patient influence toward congruity (schematically this would be $TC > PC$), and (3) data pairs showing Therapist influence toward incongruity is greater than data pairs showing Patient influence toward incongruity (schematically this would be $TI > PI$).

Yee and Gage point out that use of the frequency-of-shift-across-median technique means that not all of the data are analyzed (in their study 60 per cent were non-changers). Consequently, the authors recommend use of a second technique which is closely related to the first, but which makes use of all data. Yee and Gage call it the frequency-of-change-in-product-moment. To set up and analyze the data units for this technique, five steps are necessary.

1) Convert the data from raw scores to standard scores.
2) Compute the cross-product of each data pair's $\underline{z}$ scores for all pre-test data and then separately compute the cross-product of each data pair's $\underline{z}$ scores for all post-test data.
3) Compare cross-products. If $\underline{z}T_1\underline{z}P_1 > \underline{z}T_2\underline{z}P_2$, then the change is toward congruency. If $\underline{z}T_2\underline{z}P_2 > \underline{z}T_1\underline{z}P_1$, then the change is toward incongruency. From this one knows the direction of influence.
4) Compute the cross-lagged $\underline{z}$ products for each pair's $\underline{z}$ scores.
5) Compare the cross-lagged $\underline{z}$ products. Yee and Gage summarize the procedures by writing:

> "If change is toward congruency and
> if $\underline{z}T_1\underline{z}P_2 > \underline{z}T_2\underline{z}P_1$, then T is source of influence;
> if $\underline{z}T_1\underline{z}P_2 < \underline{z}T_2\underline{z}P_1$, then P is source of influence.
> If change is toward incongruency, and
> if $\underline{z}T_1\underline{z}P_2 > \underline{z}T_2\underline{z}P_1$, then P is source of influence;
> if $\underline{z}T_1\underline{z}P_2 < \underline{z}T_2\underline{z}P_1$, then T is source of influence.

From this step we determine the source of influence.

There are two main problems with the Yee and Gage technique. Of primary importance is the effect of comparing two variables, one of which is quite stable while the other is not. This fact alone seems to predetermine which variable will appear as the cause (the stable one) and which will appear as the effect (the unstable one). Pelz has worked through this particular problem of the FCP technique in a memo concerning an application of the Yee and Gage analysis (1968). Pelz demonstrates the artifact of the FCP which is that the "variable which is more stable through time will have a greater likelihood of appearing causally dominant." Pelz says the problem lies in the inappropriateness of the null hypothesis, which says that each category (congruent change with X influence, incongruent change with X influence, congruent change with Y influence, and incongruent change with Y influence) has an equal likelihood of being filled with cases. He suggests that a new baseline be investigated, possibly a set of expected values based on the auto correlations and the cross correlations. Then departure from the expected frequencies could be tested by chi-square.

The second problem is that the FCP method has no control for third-variable causation. For instance, let's say that a researcher measured the height and weight of a group of women at two time points. Unfortunately, the researcher is not aware that each member of the group is also a member of TOPS (Take Off Pounds Sensibly). The result would show that height caused weight. However, a third variable, lower food consumption, was actually the causal variable. Although the example may be a bit bizarre, it illustrates the technique's lack of ability to control for spurious relationships.

IV.2.D. Heise's Path Analysis

One last approach to causal inference will be reviewed before comparing the techniques. It is path analysis, originally developed by Wright (1934) and extended by Heise (1969) to encompass time-lagged data.

Path coefficients are derived by partitioning correlations into components that "represent the amount of change that will occur in a dependent variable given a one-unit change in an independent variable with all other variables held constant and measurements expressed in standardized form." (Heise, 1969). Its assumptions include linearity, constancy in the causal relationship, causal processes that apply to the entire sample, outside disturbances to the process that operate at a constant rate, measurement errors that remain constant and are uncorrelated, and causal processes that operate instantaneously rather than through a storage process.

The path coefficient is similar to partial correlation. The extent of this similarity can be seen by comparing the two coefficients:

$$r_{14.2} = \frac{r_{14} - r_{12}r_{42}}{\sqrt{(1-r_{12}^2)\,(1-r_{42}^2)}} \qquad p_{41} = \frac{r_{14} - r_{12}r_{42}}{1 - r_{12}^2}$$

Heise explains that his approach is dissimilar to the Pelz and Andrews technique of partialling in that he determines the effect of one variable (A) on the other (B) relative to the total variance in B while partial correlation gives the effect of A on B relative to the variance that occurs after the original measurement.

IV.2.E. A Comparison of the Techniques

In considering the differences between the Rozelle and Campbell model and the Yee and Gage frequency-of-change-in-product-moment model, one sees differing expectations for the synchronous correlations. Rozelle and Campbell expect equal *r*'s and Yee and Gage expect increasing or decreasing synchronous *r*'s. Perhaps an analogy will help explain these different approaches to model building. Consider, for example, that one has a loaf of bread. There are several ways he can slice it. He can either take two slices out of the middle of the loaf or he can cut off one end piece and one center piece. If he does the former (as Rozelle and Campbell do), then he expects each slice of bread to be as influenced by the flour, eggs, milk, and yeast as is the next slice. However, if he does the latter (as Yee and Gage do), then he expects the end slice to be different from the inside slice. Such is the case with the two models. Campbell says that the two synchronous correlations will be the same except for sampling or random error, because he envisions time-series data where the researcher puts his measuring tool into an ongoing process to determine causal influence. However, other data sets begin at the start of the causal relationship. For instance, Yee and Gage measure teachers and pupils very early in the school year; one could argue that they are looking at the beginning of the process (which is analogous to taking the end slice of bread). Another example that is even more heavily weighted on the side of measurement at the initial stage of a relationship is the Howard and Krause data (1969). In many instances, patients and therapists are measured after each session, including the first session. Here it would be unreasonable to assume that initial synchronous correlations would be comparable to latter synchronous correlations.

Another difference between the two techniques involves different assumptions about the variance. Each method reflects the originator's expectations. Campbell expects the variance to remain the same (hence the use of r) while Yee and Gage expect the variance to change (hence the use of cross-products).

These differing expectations in two of the methodologies lead us to realize that we need a more complete comparison of basic assumptions across all methodologies. The following table is a first attempt to make these comparisons. As it stands it is necessarily unfinished. The

Table IV.2/3

Techniques and Assumptions

	Correlation Coefficients		Data			Analysis Model	Effects	
	Synchronous	Auto	Linear-Nonlinear	Dichotomous-Continuous	Static-Dynamic	Univariate-Multivariate	Immediate-Delayed	Handle Spurious Relationships
Lazarsfeld [a]	Change from T1 to T2	Not stable	Both	Dichotomous	S-S S-D	Univariate	Delayed	Indirectly[b]
Rozelle and Campbell	No change	Stable	Linear	Continuous	S-S S-D[f]	Univariate	Delayed	Yes[c]
Yee and Gage (FSM)	Change from T1 to T2	Not stable	Both	Trichotomous	S-S S-D	Univariate	Delayed	No
Yee and Gage (FCP)	Change from T1 to T2	Not stable	Linear	Continuous	S-S S-D	Univariate	Delayed	No
Heise[d]	No change	Stable	Linear	Continuous	S-S S-D[f]	Multivariate	Immediate	No[e]

a. Technique is same for 1948 and 1969, only method of evaluating sixteen-fold table has changed.

b. By repeating analysis on several possible relationships.

c. To compute the no-cause baseline one needs reliability data, and one must be able to meet assumptions for the synchronous and auto correlations.

d. Heise has systematically relaxed many of the assumptions and noted the effects (1969).

e. Assumes all relevant variables already in the analysis.

f. If three time points instead of two, then also D-S and D-D relationships.

columns (the assumptions) are not exhaustive, but represent a present statement of the problem. The rows (the techniques) are not exhaustive. Pelz's approach might be added as well as others working in the area (e.g. Duncan and Blalock). However, given these limitations, Table IV.2/3 still serves as an organizational framework for causal inference models.

From a consideration both of the techniques in general and of the specifics of the table, two issues seem to need attention. First, there is the problem of what to do about possible spurious relationships. The second, and related problem, is a general inability to incorporate multivariate procedures in the data analysis. Each of these deserves a separate discussion.

IV.2.F. Spurious Relationships

Very often one has a data set that includes several variables. This is especially true for panel or repeated measures studies. Yet each of the present techniques, with the possible exception of Heise's, uses only two variables. It seems quite likely in many data sets that other measured variables impinge on the relationship of the given pair. When a researcher is engaged in an explanatory study, he finds that it is not enough to say that A is the cause of B. He must also be aware of other variables that might be the real causes or might combine with A jointly to cause B or specify the situation or subpopulations in which A causes B.

Borrowing from Hyman (1955) we note three ways in which relationships are elaborated:

1. Explanation: relationship is spurious; a third, antecedent, variable is the real cause.

$$\begin{tikzcd} & X \arrow[dl] \arrow[dr] & \\ A & & B \end{tikzcd}$$

2. Interpretation: the relationship occurs through a third variable which intervenes between A and B.

$$A \longrightarrow X \longrightarrow B$$

3. Specification: a third variable extablishes conditions in which the original relationship is stronger or weaker.

$$\overset{X_1}{A \longrightarrow B} \qquad\qquad \overset{X_2}{A \longrightarrow B}$$

The following steps suggest analytical methods for handling each of the three types of elaboration:

1. Use path analysis or partial correlation to handle the third variable as explanatory.
2. Use path analysis or part correlation to handle the third variable as interpretive.
3. Use a difference in *r* (not preferred because involves a dichotomization of *X*) or path analysis (but still have problem with *X*) for third-variable specification.

At this time, all we can say is that we are aware of the third-variable problem and see that it is inappropriate to consider two variables in isolation. However, we lack strong ways of handling the third variable. Other solutions will become apparent later. We are using our prerogative, as researchers, to ask questions even when the answers are not close at hand.

IV.2.G. Data Analysis[1]

Before any causal technique can be used, we must first know the nature of the data. This phase is described as "data inspection in the pre-analysis stage". Then a possible multivariate technique that can handle several variables is outlined in "data procedures in the mid-analysis stage". A final section called "data procedures in the post-analysis stage" suggests procedures that are appropriate for the elaboration of relationships which have emerged in the mid-analysis stage.

IV.2.H. Data Inspection in the Pre-Analysis Stage

The pre-analysis stage of a non-experimental causal inference study begins when there are coded data from two time points and ends when the data are ready for correlation analysis. Four steps are suggested:

1. Histograms. The effect of skewness on correlation points to the need to examine the distribution of each variable in the analysis. The histogram displays the extent of univariate normality and provides necessary evidence for deciding on possible transformations.

---

[1] While this manuscript was in preparation, M. Paisley tested the analysis steps that appear as suggestions in IV.2.G. - IV.2.J. The interested reader is urged to consult her dissertation, (M. Rees) A comparison of cross-lagged, path, and multivariate causal inference techniques applied to interest, information, and aspiration among high school students, Department of Psychology, Northwestern University, 1971.

2. Transformations. When skewness exists, distributions are made more symmetrical by transformation of scores. The effectiveness of the transformation is checked by comparing correlations of transformed variables with those of original variables. A successful transformation will increase the correlation.
3. Scatter Plots. Bivariate linearity is one assumption of product moment correlation. Scatter plots, by displaying the joint distribution of two causal variables, give the information needed for a decision to compute etas.
4. Etas. When scatter plots indicate that bivariate nonlinearity exists, etas are computed. Eta is based on variance around array means in a one-way analysis of variance design in which one of the two variables has been reduced to categorical status (thus providing discrete arrays) while the other remains continuous. This results in a coefficient that is higher than the product moment coefficient (which is based on variance around the regression line) when the pattern of means is curvilinear. Difference in variance accounted for by the regression line and by array means can be tested for statistical significance.

IV.2.I. Data Procedures in the Mid-Analysis Stage

The mid-analysis stage begins when data have been prepared for correlation analysis and ends when the lagged causal relationships have been computed. Five steps are suggested:

1. Factor analyze variables. The problem of independence calls for a first step of factorization. Both the unrotated and the rotated solutions provide useful information. The unrotated solution is the check for a common factor. Although it is anticipated that many variables will initially load on this factor, the amount of variance accounted for is the crucial test. If the first factor represents less than 51% of the variance, the assumption is made that variables are multi-factored. Next the rotated solution is examined. Variables that load strongly on a factor are combined into indices.
2. Construct Indices. The factor analysis results in a decision to combine certain variables -- each set having focused a factor. Addition of z scores of each set provides an adequate way to compute the index. It should be remembered that two kinds of indices need to be computed. The causal construct index is computed separately for each year. However, a control variable index, if the variable set has been measured at two time points, is an average of the two.
3. Extract Third-Variable Covariance. Each set of four causal constructs (CC_1 and CC_2 at T_1 and T_2) alternately serve as Y', predicted by the equation:

$$Y' = b_1CV_1 + b_2CV_2 + b_nCV_n.$$

This procedure extracts the amount of variance shared by a CC (causal construct) and CVs (control variables).

Both endogenous and exogenous variables enter the analysis at this time. However, exogenous variables have the status only of control variables, while endogenous variables can be either control variables or causal variables. (It is possible that a construct would be used both ways during the analysis of a data set.)

4. Compute Residual Scores. Each four sets of scores to be correlated are the residuals computed from:

$$\text{Residual} = Y_i - Y'_i$$

5. Correlate Residuals. The six correlations among the four residual scores are computed. Since third variable covariance has been extracted from the synchronous correlations, the lags are compared with each other and against the synchronous correlation.

IV.2.J. Data Procedures in the Post-Analysis Stage

The post-analysis stage begins when six correlations have Been computed between each time-lagged pair and ends when relationships have been elaborated as much as the data permit. Three steps are suggested:

1. Compute partial correlations. Although third-variable covariance has been removed, synchronous correlations still contain variance shared by the two causal constructs. Partialing on "the other causal construct" removes the variance it shares with the two variables in the lag. The partial correlation of CC_1 T_1 with

CC_2 T_2 has the covariance of CC_2 T_1 removed. Similarly, the CC_2 T_1 - CC_1 T_2 relationship, after partialing, has the effect of CC_1 T_1 removed.

2. Partition for Specification. The covariance extraction technique results in a single set of correlations between variables from which effects of certain third variables have been removed. While the technique permits explanation of a relationship or the discovery of a hidden relationship, it does not allow for specification of a relationship. To achieve specification, the control variable is split at its median and correlations are computed for each subsample. Then the relationship between the constructs in the subsample scoring low on the control variable and in the subsample scoring high on the control variable can be examined. In this way, differential effects are brought to light.

3. Compute Multiple Correlation. As causal inference techniques, such as this one, open the way for consideration of more than two variables or constructs in a data set an additional procedure is called for. What is needed is a procedure that allows one to see the causal relationship over time when two (or more) constructs combine to cause the third. For instance, if factor analysis shows the separateness of three variable sets that seem to form a causal system, one would first analyze them two at a time. Results, for example, might show that CC_1 influences CC_2 and also that CC_3 influences CC_2. At this point, multiple correlation

could be employed to show the extent to which CC_1 and CC_3 combine to cause variation in CC_2.

IV.2.K. Overview

The reader may feel by now that causal inference is so deeply mired in methodological problems that it will never be on dry land again. However, it is the intent of this chapter to say that, while we cannot solve all the problems, we can arrive at a solution that is better than the present state. This reasoning follows Simon's (1957) notion of the optimizing and satisficing functions of science.

Our present knowledge of causal inference has been gained in this way. Lazarsfeld did not solve all problems, but advanced our thinking to two-time-point analysis. Campbell and Stanley did not solve all problems, but introduced continuous variables that are measured at two time points. Rozelle and Campbell made us aware of the greater number of competing causal hypotheses. We have similar debts to the other researchers cited in this section. In recent years, efforts to strengthen causal inference have become unusually collaborative and collegial. We believe that the analyses we are undertaking, using several data sets within the framework of the convergent-discriminant matrix, will contribute to the collective enterprise.

V. ON THE CORRELATION COEFFICIENT

Section IV's discussion of correlational causal inference prompts us to return to the question of correlational inference in general. Section II discussed some technical considerations in computing correlations across aggregate data. The import of Section II is that aggregate correlations are *prima facie* neither valid nor invalid. Validity depends on the nature of the data and the care with which aggregate units are handled.

There is a literature on the robustness of the Pearson product-moment correlation coefficient. Carroll (1961) and Paisley (1965) exemplify studies in which the assumptions underlying Pearson *r* are violated to determine the seriousness of the correlational underestimate in each case. These and other studies confirm that Pearson *r* is a robust statistic except in the face of extreme skewness, range attenuation, interval degredation, etc. When extreme problems exist, the researcher can shift to another procedure, such as rank correlation. If factors mitigate against other procedures, the researcher can adopt an "only wheel in town" philosophy -- use Pearson *r*, but heuristically, in anticipation of better sequel data.

Correlation versus individual fluctuation. The use of correlation coefficients is sometimes criticized on a non-statistical level. For example, Lazarsfeld (1948, 1969) rejects correlational causal inference even when the assumptions of Pearson *r* are fully met. He fears that the behavior of individual cases will be averaged in a way that masks "turnover" or "migration" of those cases from one behavioral state to another (e.g., from an intention to vote for Willkie to an intention to vote for Roosevelt).

It is possible to construct tables in which the correlation coefficient disagrees with the apparent "turnover" trend. Therefore one must accept Lazarsfeld's contention that a correlation coefficient is not universally valid.

However, arguments based on exceptional cases overlook the probabilistic nature of all statistics. In computing correlations to fill a convergent-discriminant matrix, we commit Type I and Type II errors a certain percentage of the time. Assuming a hypothetical "true" relationship between variables, correlations will overstate the relationship in certain cases and understate it in others. Correlations may even state a positive relationship when the "true" relationship is negative and vice versa. Such fallibility is in itself one of the best arguments for extending multiple operationism beyond method variation into the realm of analysis models.

Correlations indeed tell us nothing about the behavior of individual cases. Analysis of residuals helps in that direction, but such analysis tends to be _post_ _hoc_ and particularistic.

Correlations are valuable _because_ of their averaging function, not in spite of it. A correlation coefficient tells us the overall correspondence between distributions of cases. Instead of reporting the position of any single case on any distribution, a correlation coefficient tells us the average extent to which the positions of all cases _match_ on two (or more) distributions.

It is parsimonious to know patterns of phenomena rather than individual fluctuations. The usefulness of correlation in social research lies in its power to show pattern correspondence between distributions of cases.

The behavior of individual cases within those distributions is another kind of data that can generate insight on exceptionality. However, we rely on correlations to state the overall pattern that provides a context for interpreting exceptional cases.

VI. SOME INDICATED DIRECTIONS FOR RESEARCH

Looking ahead, there are two related questions: (1) What can research tell us about the logic of the convergent-discriminant matrix? (2) What can the logic of the convergent-discriminant matrix tell us about research?

Data runs will help to answer the first question. The unique roles of each matrix, used with a particular replication dimension, will be clearer. The Section III of a later draft of this paper will probably be organized in terms of functional differences rather than structural differences.

On the question of the CDM's contribution to research, we would like to see the entire set of convergent-discriminant matrices used sequentially throughout the history of a research project. That is, remembering that two dimensions (measures and analysis models) are methodological in character while the other three (concepts, populations, and times) are substantive, we would like to see the one purely methodological matrix used in early research phases, followed in later phases by methodological-substantive matrices and finally, at the moment of inference, by the purely substantive matrices. One sequence would be:

1. The one purely methodological matrix, measures by analysis models, alerts the researcher to data anomalies at the statistical level. The analyzability

of measures ought to be considered even before their final form is established.

2. Then analysis proceeds to methodological-substantive comparisons. The multitrait-multimethod matrix, concepts by measures, establishes construct validity.
3. Measures by populations analysis shows concern for the stability of a measure across populations.
4. The measures by times matrix serves a similar purpose, particularly if instrument decay is suspected.
5. The concepts by analysis models matrix leads to the preference of one analysis model over its alternatives. The CA matrix is redundant, however, if the concept has been operationalized in terms of one measure only and if that MA matrix has already been studied.
6. The chosen analysis model can next be tested for robustness in a populations by analysis model matrix. The PA analysis would most be needed in the case of a temperamental statistic (e.g., stepwise regression) or a postulated but suspect data-fitting (e.g., fitting an exponential diffusion function in information-flow research).
7. The times by analysis models matrix serves a similar purpose for the time dimension.
8. If the research project has survived these checks and cross-checks, then it can proceed into its inferential phase and the three purely substantive matrices. The concepts by populations matrix supports an inference of external validity.

9. The concepts by times matrix may support an inference of causal direction between concepts.
10. The tenth and last matrix, populations by times, may support an inference that one population or person is becoming more similar to -- modelling -- another population or person over time.

Of course, no research project can thus be reviewed with CDMs unless conceptualization, design, data collection, and analysis are rich on all five dimensions. Someday a design may be considered incomplete if it omits important dimensions of variation, but now it is likely that the time dimension and/or the population dimension will be missing. Single operationism is still the order of the day: one concept, one measure, one population, one time, one analysis model.

We hope that behavioral research will move from single operationism to "deliberate variation at every research choice-point where a reliability or validity challenge exists." The first consequence will be a great proliferation of data. It may seem that the CDM adds to this problem. However, convergence and discrimination are stringent tests. The CDM convicts more than it acquits.

Multiple operationism leads to fat research projects; the task of the CDM is to make them trim again.

REFERENCES

Blalock, H.M. _Causal inference in nonexperimental research._ Chapel Hill: U. North Carolina Press, 1961.

Blalock, H.M. & Ann B. Blalock. _Methodology in social research._ New York: McGraw-Hill, 1968.

Bock, R.D. & R.E. Bargmann. Analysis of covariance structures. _Psychol. Bull._, 1966, _31_, 507-534.

Borgatta, E.F. The stability of interpersonal judgments in independent situations. _J. abn. soc. psychol._, 1960, _60_, 188-194.

Borgatta, E.F. The structure of personality characteristics. _Behavioral science_, 1964, _9_, 8-17.

Borgatta, E.F. A very short test of personality: the behavioral self-rating form. _Psychol. reports_, 1964, _14_, 275-284.

Borgatta, E.F. & G.W. Bohrnstedt. _Sociological methodology 1969._ San Francisco: Jossey-Bass, 1969.

Boruch, R.F. _et al_. Alternative methods of analysis: multitrait-multimethod data. _Educ. psychol. measurement_, Winter 1970.

Boruch, R.F. & L. Wolins. _A procedure for estimating the amount of trait, method, and error variance attributable to a measure._ Paper presented at the Psychometric Society Convention, University of North Carolina, 1968.

Boudon, R. A new look at correlational analysis. In: Blalock & Blalock, 1968.

Brewer, M.B., D.T. Campbell & W.J. Crano. Testing a single-factor model as an alternative to misuse of partial correlation in hypothesis-testing research. _Sociometry_, 1970.

Campbell, D.T. Definitional versus multiple operationism. _et al_, 1969, _2_, 14-17.

Campbell, D.T. From description to experimentation. In: Harris, 1963.

Campbell, D.T. & D.N. Clayton. Avoiding regression effects in panel studies of communication impact. _Studies in public communication_, 1961, _3_, 99-118.

Campbell, D.T. & D.W. Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol. bull., 1959, 56, 81-105.

Campbell, D.T. & E.J. O'Connell. Methods factors in multitrait-multimethod matrices: multiplicative rather than additive? Multivariate beh. res., 1967, 2, 409-426.

Campbell, D.T. & J.S. Stanley. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.

Carroll, J.B. The nature of the data, or how to choose a correlation coefficient. Psychometrika, 1961, 26, 347-372.

Cartwright, D. Ecological variables. In: Borgatta & Bohrnstedt, 1969.

Cattell, R.B. Handbook of multivariate experimental psychology. Chicago: Rand McNally, 1966. *

Coleman, J.S. Introduction to mathematical sociology. Glencoe, Ill.: Free Press, 1964.

Cronbach, L.J. The two disciplines of scientific psychology. Amer. psychol., 1957, 12, 671-684.

Cronbach, L.J. & G.C. Gleser. Assessing similarity between profiles. Psychol. bull., 1953, 6, 456-473.

Cronbach, L.J. & P.E. Meehl. Construct validity in psychological tests. Psychol. bull., 1955, 52, 281-302.

Duncan, O.D. & B. Davis. An alternative to ecological correlation. Amer. soc. rev., 1953, 18, 665-666.

Fanshel, D., L. Hylton & E.F. Borgatta. A study of behavioral disorders in children in residential treatment centers. J. psychol. studies, 1963, 14, 1-23.

Feigl, H. & May Brodbeck. Readings in the philosophy of science. New York: Appleton-Century-Crofts, 1953.

Fruchter, B. Introduction to factor analysis. New York: Van Nostrand, 1954.

Goodman, L.A. Ecological regressions and the behavior of individuals. Amer. soc. rev., 1953, 18, 663-664.

Goodman, L.A. Some alternatives to ecological correlation. Amer. j. soc., 1959, 44, 610-625.

*Cattell, R.B. The three basic factor-analytic research designs - their interrelations and derivatives. Psychol. bull., 1952, 49, 499-520.

Harris, C. Problems in measuring change. Madison: U. Wisconsin Press, 1963.

Heise, D.R. A model for causal inference from panel data. Madison: Department of Sociology, U. Wisconsin, 1969.

Heise, D.R. Problems in path analysis and causal inference. In: Borgatta & Bohrnstedt, 1969.

Hempel, C.G. Fundamentals of concept formation in the empirical sciences. Int. encyclopedia of unified science, II, 7. Chicago: U. Chicago Press, 1952.

Hempel, C.G. The philosophy of natural science. Englewood Cliffs, N.J.: Prentice Hall, 1966.

Hornik, R. Review of selected nonexperimental causal analysis techniques. Stanford: Institute for Communication Research, Stanford U., 1969.

Howard, K.I. & M.S. Krause. Some comments on "Techniques for estimating the source and direction of influence in panel data". Psychol. bull., 1970.

Hyman, H. Survey design and analysis. Glencoe, Ill.: Free Press, 1955.

Jackson, D.N. Multimethod factor analysis in the evaluation of convergent and discrimiant validity. Psychol. bull., 1969, 72, 30-49.

Kaiser, H.F. Relating factors between studies based upon different individuals. Urbana: Bureau of Educ. Research, U. Illinois, 1960.

Kenny, D.A. A refined no-cause baseline for the cross-lagged panel correlation. Evanston: Department of Psychology, Northwestern U., 1969.

Kerlinger, F.N. Foundations of behavioral research. New York: Holt, Rinehart, Winston, 1964.

Lazarsfeld, P.F. A memoir in honor of Professor Wold. New York: Department of Sociology, Columbia U., 1969.

Lazarsfeld, P.F. Mutual effects of statistical variables. New York: Department of Sociology, Columbia U., undated (ca. 1948).

Lazarsfeld, P.F. & A.H. Barton. Qualitative measurement in the social sciences: typologies, classification, and indices. In: Lerner & Lasswell, 1951.

Lazarsfeld, P.F. Problems in the analysis of mutual interaction between two variables. Memorandum to the NBC Research Department. New York: Department of Sociology, Columbia U., 1969.

Lerner, D. & H.D. Lasswell. The policy sciences. Stanford: Stanford U. Press, 1951.

McNemar, Q. Psychological statistics. New York: Wiley, 1962.

Maher, B.A., N. Watt & D.T. Campbell. Comparative validity of two projective and two structured tests in a prison population. J. appl. psychol., 1960, 44, 284-288.

Moran, L.J. Generality of word-association response sets. Psychol. mono., 1966, 80, 1-25.

Nagel, E. The structure of science. New York: Harcourt, Brace & World, 1961.

Norman, W.T. Relative importance of test item content. J. consulting psychol., 1963, 27, 166-174.

Norman, W.T. Toward an adequate taxonomy of personality attributes: replicated factor structure in peer nomination personality ratings. J. abn. soc. psychol., 1963, 66, 574-583.

Nunnally, J. The analysis of profile data. Psychol. bull., 1962, 59, 311-319.

Paisley, W.J. Correlational analysis and "The nature of the data". Stanford: Institute for Communication Research, Stanford U., 1965.

Paisley, W.J. Some decision points in the survey research process. In: Shaw, 1967.

Paisley, W.J. & L. Ruggels. Choosing a correlation coefficient for skewed (communication development) data: an empirical comparison of five measures. Stanford: Institute for Communication Research, Stanford U., 1966.

Parker, E.B., D.A. Lingwood & W.J. Paisley. Communication and research productivity in an interdisciplinary behavioral science research area. Stanford: Institute for Communication Research, Stanford U., 1968.

Pelz, D.C. & F.M. Andrews. Detecting causal priorities in panel study data. Amer. soc. rev., 1964, 29, 836-848.

Popper, K.R. The logic of scientific discovery. New York: Basic Books, 1961.

Ranney, A. Essays on the behavioral study of politics. Urbana: U. Illinois Press, 1962.

Ranney, A. The utility and limitations of aggregate data in the study of electoral behavior. In: Ranney, 1962.

Rees, M.B. A comparison of cross-lagged, path, and multivariate causal inference techniques applied to interest, information, and aspiration among high school students. Evanston: Department of Psychology, Northwestern University. (Ph.D. dissertation)

Rees, M.B. & D.T. Campbell. Linguistic and structural effects in attitude scales. Research in progress, Department of Psychology, Northwestern University.

Robinson, W.S. Ecological correlations and the behavior of individuals. Amer. soc. rev., 1950, 15, 351-357.

Rozelle, R.M. & D.T. Campbell. More plausible rival hypotheses in the cross-lagged panel correlation technique. Psychol. bull., 1969, 71, 74-80.

Rubin, E.L. Structural analysis of the multitrait-multimethod matrix. Evanston: Department of Psychology, Northwestern U., 1970. (Dissertation proposal)

Russett, B.M. et al. World handbook of political and social indicators. New Haven: Yale U. Press, 1964.

Schuessler, K. Causality versus communality. et al, 1969, 2, 21-23.

Shaw, D.L. Decision points in mass communications research: survey, content analysis, historical, and experimental methods. Studies in journalism and communication, 6. Chapel Hill: School of Journalism, U. North Carolina, 1967.

Simon, H.A. Models of man. New York: Wiley, 1957.

Stephenson, W. The study of behavior: Q-technique and its methodology. Chicago: U. Chicago Press, 1953.

Taylor, C.L. Aggregate data analysis: political and social indicators in cross-national research. Paris: Mouton, 1968.

Tucker, L.R. Some mathematical notes on three-mode factor analysis. Psychometrika, 1966, 31, 279-311.

Wallace, J. & L. Sechrest. Frequency hypotheses and content analysis of projective techniques. J. consulting psychol., 1963, 27, 387-393.

Webb, E.J. et al. Unobtrusive measures: nonreactive measurement in the social sciences. Chicago: Rand McNally, 1966.

Webster, H. & C. Bereiter. The reliability of changes measured by mental test scores. In: Harris, 1963.

Wright, S. The method of path coefficients. Annals of mathematical statistics, 1934, 5, 161-215.

Yee, A.H. & N.L. Gage. Techniques for estimating the source and direction of causal influence in panel data. Psychol. bull., 1968, 70, 115-126.

Zuckerman, M. et al. A multitrait-multimethod approach to the traits (or states) of anxiety, depression and hostility. J. proj. tech. & personality assessment, 1967, 31, 39-48.